



UNIVERSITÀ
POLITECNICA
DELLE MARCHE

Università Politecnica delle Marche
Scuola di dottorato di ricerca in scienze dell'ingegneria
curriculum in ingegneria informatica gestionale e automazione

DL4US: Unlocking the potential of deep learning for ultrasound image analysis in gynecology and rheumatology

Candidate:
Maria Chiara Fiorentino

Advisor:
Prof. Emanuele Frontoni

Coadvisor:
Dr. Sara Moccia

Academic Year 2022-2023



UNIVERSITÀ
POLITECNICA
DELLE MARCHE

Università Politecnica delle Marche

Scuola di dottorato di ricerca in scienze dell'ingegneria
curriculum in ingegneria informatica gestionale e automazione

DL4US: Unlocking the potential of deep learning for ultrasound image analysis in gynecology and rheumatology

Candidate:

Maria Chiara Fiorentino

Advisor:

Prof. Emanuele Frontoni

Coadvisor:

Dr. Sara Moccia

Academic Year 2022-2023

Abstract

Medical imaging is important to healthcare because it can be used for early illness diagnosis, staging, and management, as well as for choosing the most appropriate treatment and monitoring its course of action. Ultrasound imaging stands out among other imaging diagnostic mediums due to its real-time nature as a highly cost-effective tool that offers the practitioner an unmatched level of involvement. The presence of speckles and acoustic noise, which make it difficult to interpret images, the high reliance on diagnostic operator experience, and the high inter- and intra-observer variability across different institutes and manufacturers are all unique obstacles that US systems have to overcome. In order to make US diagnoses and assessments more accurate and reliable, a series of image analysis methods have been proposed over the year. By extrapolating hidden patterns from data and exploiting its excellent ability to make decision rules, Deep Learning (DL) has set itself to become one of the most important tools in areas such as image recognition and computer vision. The overall goal of this PhD thesis is to exploit DL algorithms for US image analysis. In particular, this work focuses on two specific fields: i) fetal US imaging, in which US is used daily in clinical practice and has been extensively researched, including preexisting DL applications and ii) rheumatology, where DL usage is quite innovative and, thus, there is still much to be done in relation to DL methodologies and applications. In particular, the contributions of this PhD work are:

- *introducing and discussing the most innovative and effective DL methods found in the literature for fetal US image analysis. An updated review may be a valuable and compact source of information for young researchers and a reference document for those already working in the field which may develop innovative fetal architectures*
- *Developing a new method which exploits a distance field strategy to precisely extrapolate Head Circumference measurements in fetal US images*
- *Developing a new method which exploits a self contrastive learning to train a semi-supervised fetal standard plane classifier*
- *Developing a new framework for metacarpal-head cartilage-thickness estimation in US rheumatological images*

Each contribution examines a particular aspect of US image analysis field with the betterment of the patients' quality of life in mind, by ensuring a more US dependable

decision-making process while addressing some of the US open problems.

Abstract

L'imaging medico è importante per l'assistenza sanitaria perché può essere utilizzato per la diagnosi precoce, la stadiazione e la gestione della malattia, nonché per scegliere il trattamento più appropriato e monitorarne il corso d'azione. L'imaging ecografico si distingue tra gli altri mezzi diagnostici di imaging grazie alla sua natura che permette l'analisi, permettendo al professionista un livello di coinvolgimento senza pari. Sebbene vantaggiosi, la presenza di macchie e rumore acustico che rendono difficile l'interpretazione delle immagini, l'elevata dipendenza dall'esperienza dell'operatore e l'elevata variabilità inter e intra-osservatore tra diversi istituti e produttori, sono però tutti limiti che ostacolano l'utilizzo degli ultrasuoni. Al fine di rendere le diagnosi e le valutazioni ecografiche più accurate e affidabili, nel corso degli anni sono state proposte una serie di metodi di analisi automatica delle immagini. Estrapolando pattern nascosti dai dati e sfruttando la sua eccellente capacità di stabilire regole decisionali, il Deep Learning (DL) si è prefisso di diventare uno degli strumenti più importanti in aree come il riconoscimento delle immagini e la visione artificiale. L'obiettivo generale di questa tesi di dottorato è quindi sfruttare gli algoritmi DL per l'analisi delle immagini ecografiche. In particolare, questo lavoro si concentra su due campi specifici: i) l'ecografia fetale, che viene utilizzata quotidianamente nella pratica clinica e vanta comunque una estensiva ricerca nel settore, comprese le applicazioni preesistenti di DL e ii) reumatologia, in cui l'uso di DL è piuttosto innovativo e, quindi, c'è ancora molto da fare in relazione alle metodologie e alle applicazioni del DL. In particolare, i contributi di questo lavoro di dottorato sono:

- *presentare e discutere le metodiche DL più innovative ed efficaci presenti in letteratura per l'analisi delle immagini ecografiche fetali. Una revisione aggiornata può essere una preziosa e compatta fonte di informazioni per i giovani ricercatori e un documento di riferimento per coloro che già lavorano nel settore che vogliono sviluppare architetture fetali innovative*
- *Sviluppo di un nuovo metodo che sfrutta una strategia di regressione del contorno per estrapolare con precisione la misura della circonferenza della testa in immagini ecografiche*
- *Sviluppo di un nuovo metodo che sfrutta un apprendimento self-contrastativo per addestrare un classificatore capaci di distinguere vari piani standard fetali*

- *Sviluppo di un nuovo framework per la stima dello spessore della cartilagine della testa metacarpale in immagini reumatologiche*

Ogni contributo esamina un aspetto particolare del campo dell'analisi delle immagini ultrasuoni con in mente il miglioramento della qualità della vita dei pazienti, garantendo un processo decisionale più affidabile e veloce.

Contents

1	Introduction	1
1.1	Problem statement and motivation	1
1.2	Introduction to deep learning for ultrasound image analysis	2
1.3	Aim of the thesis	3
1.4	Structure of the thesis	4
1.5	Scientific publications	4
1.5.1	Journal publications	5
2	Literature review on fetal-ultrasound analysis	7
2.1	Introduction	7
2.1.1	Survey strategy	9
2.2	Metrics and datasets	10
2.2.1	Performance metrics	10
2.2.2	Publicly-available datasets	14
2.3	Fetal standard-plane detection	17
2.3.1	Limitations and open issues for fetal standard-plane detection	22
2.4	Anatomical structure analysis	28
2.4.1	Heart	29
2.4.2	Brain	30
2.4.3	Placenta & amniotic fluid	32
2.4.4	Others	34
2.4.5	Multi-organ analysis	37
2.4.6	Limitations and open issues for anatomical structure analysis	41
2.5	Biometry parameter estimation	43
2.5.1	Limitations and open issues for biometry parameter estimation	46
2.6	Other tasks	46
2.6.1	Gender identification	46
2.6.2	Pose estimation	47
2.6.3	Preterm-birth prediction	47
2.6.4	US simulatio	47
2.6.5	Video summarization	48
2.6.6	Shadow artifact removal	48
2.6.7	Probe movement control	49
2.6.8	Volume reconstruction	49
2.6.9	Limitations and open issues for other tasks	49

Contents

2.7	Discussion and conclusions	50
3	Automated measurement of fetal head circumference	57
3.1	Introduction	57
3.2	2step RCNN	59
3.2.1	Methods	59
3.2.2	Experimental setup	64
3.2.3	Results	66
3.3	Mask-R ² CNN	67
3.3.1	Method	68
3.3.2	Mask-R ² CNN heads	70
3.3.3	Mask-R ² CNN training and ellipse fitting	71
3.3.4	Experimental setup	71
3.3.5	Ablation study and comparison with the literature	72
3.4	Results	74
3.5	Discussion	74
3.6	Conclusions	76
4	On the use of contrastive learning for standard-plane classification	77
4.1	Introduction	78
4.2	Related work	79
4.3	Method	80
4.4	Results and discussion	84
4.5	Conclusion & Future Directions	87
5	A deep-learning framework for metacarpal-head cartilage-thickness estimation	89
5.1	Introduction	89
5.1.1	Related work	91
5.1.2	Contribution	92
5.2	Methods	93
5.2.1	Data preparation	94
5.2.2	Cartilage-surface distance-field CNN	96
5.3	Experimental setup	97
5.3.1	Dataset	97
5.3.2	Parameter setting	98
5.3.3	Performance metrics	101
5.3.4	Ablation study & intra-clinician variability assessment	101
5.4	Results	102
5.5	Discussions	103
5.6	Conclusions	105

6	Conclusions	107
6.1	Thesis Contributions	107 6.2
	Impact	110
6.2.1	Future Perspective	111

List of Figures

1.1	Illustration of medical US analysis surveyed by [1]	2
1.2	Graphical abstract of this PhD thesis. The aim is to develop automatic algorithms which may help clinicians having more reliable and robust US image analysis. In particular, Chapter 3 and 4 present fetal US DL algorithms to delineate HC and perform scan plane classification, respectively. In Chapter 5 a DL algorithm for cartilage thickness estimation is described.	4
2.1	Summary of the tasks surveyed.	8
2.2	Summary of the papers included in this work considering: (a) addressed task, (b) year of publication, (c) contribution (journal/ conference), (d) percentile according to CiteScore rank 2020 (only for journal papers), SPD = standard-plane detection, ASA = anatomical structure analysis, BPM = biometry parameter estimation.	10
2.3	Visual samples of the most common fetal standard planes.	16
2.4	Most common deep-learning strategies to fetal standard-plane detection. FASP = fetal abdomen standard plane, FBSP = fetal brain standard plane, FFSP = fetal facial standard plane, FFESP = fetal femur standard plane.	20
2.5	Overview of the most common tasks in the literature for fetal organs analysis.	28
2.6	Most addressed tasks (from left to right) in the field of anatomical structure analysis. Lilac = heart, cyan = brain, purple = placenta and amniotic fluid, yellow = multi-organ, blue = others, temp = temporal information.	36
2.7	Number of papers surveyed in Sec. 2.4 according to the anatomical structure.	38
2.8	Common biometry parameters. Transverse section of: (a) femur length (FL), (b) abdominal circumference (AC), (c) head circumference (HC), biparietal diameter (BPD) and occipito-frontal diameter (OFD), (d) trans-cerebellar diameter (TCD).	42
2.9	Most common deep-learning strategies to fetal HC measurement.	42
2.10	Overall workflow of a computer-assisted tool to support fetal clinicians.	50
3.1	Samples of ultrasound images acquired at the (a) first and (b) second trimester. The head-circumference annotations in shown in light blue.	58

List of Figures

3.2	Workflow of the proposed framework for head-circumference delineation.	58
3.3	Fetal-head localization and centering: from the predicted bounding box (pink rectangle), the coordinates of the head-circumference (HC) centre are estimated. The centered image (light-blue rectangle) is obtained by centering the original image with respect to the HC center. The centered image has dimension $n \times m$ (with n and m equal to 705 and 545, respectively) to match the biggest bounding box found in the training set. Zero padding is added when the HC is too close to image borders.	60
3.4	Ground-truth generation for the regression network. The raw image in (a) is annotated by drawing the ellipse representing the head circumference (HC) showed in light blue in (b). The distance field is then built following a Gaussian profile (thickness r pixels), as shown in (c), to obtain the training mask showed in (d).	60
3.5	UNet architecture. The architecture includes four contraction blocks along with a bottleneck section and four expansion blocks, as described in Sec. 3.2.1.	60
3.6	In (a) and (c), the regression output is shown for two test images. The corresponding distance distributions are shown in (b) and (d), respectively, with the corresponding boxplot and outliers on top. . .	61
3.7	Head-circumference (HC) parameters obtained from the fitted ellipse: (x_c, y_c) = HC center coordinates, a = semi-major axis length, b = semi-minor axis length, and ϑ = angle of orientation.	62
3.8	Challenges in the testing datasets include different fetus head dimensions (a, b, h) and position (d, f) in the image, partially visible head skull (c), presence of noise and US image artifacts (e, g).	63
3.9	Visual samples of the ablation-study results. Each colored panel corresponds to a test image. Each panel shows the results obtained by (1st column) shallow regression, (2nd column) short proposed UNet, (3rd column) regression only and (4th column) proposed framework. Each column shows the (first row) prediction, (second row) post processed and (third-row) ellipse-fitted image.	65
3.10	Mask-R ² CNN schematic architecture. Mask-R ² CNN predicts the head-circumference distance field. The relative bounding box is shown for visualization purposes. RPN: region proposal network, FC: fully connected layers, Conv: convolutional layers.	68
3.11	Mask-R ² CNN heads. Conv: convolution, Up conv: up-sampling + convolution. Red: convolution specification (kernel dimension, number of filters), black: feature-map size.	70

3.12	Visual samples of distance-field prediction. First row: ultrasound images, second row: prediction by [2], third row: prediction by MaskR ² CNN. MaskR ² CNN does not produce spurious predictions in challenging test images, avoiding the need of post-processing. The last row shows visual samples of fetal-head delineation with Mask-R ² CNN.	72
3.13	Visual samples of the predicted distance field overlapped on a test US image for each of the ablation study. A colormap is used for the predicted distance field for visualization purposes. First row (from left to right): Transp1, Transp2, Transp3. Second row (from left to right): Up-conv1, Up-conv2, Mask-R ² CNN.	73
4.1	Main fetal standard planes are shown as well as the main different anatomical landmarks among brain standard planes. Cb = cerebellum, Th = thalami.	78
4.2	Visual representation of E3 (left) and E4 (right) experiments described in Sec. 4.3. E2 experiment follows what is currently done in E3 without ImageNet weights initialization.	83
4.3	F1-score obtained with E1, E2 and E3 - Z-multiplane considering different percentage of training data. Results are shown for (a) fetal brain, (b) fetal thorax, (c) maternal cervix and (d) fetal femur. . .	83
4.4	F1-score obtained for E1, E2, E3 and E4 for Z-brain.	84
4.5	Confusion matrices for E4 and Z-brain considering the (a) 10%, (b) 50% and (c) 99% of training patients.	85
4.6	Visual samples of CAMs obtained from E1, E2, E3, E4 experiments for different percentage of annotated patients. Block a), b) and c) show results for an image of class TC, TT and TV, respectively. Numeric values provide the probability gained by the true class (blue labels), whereas red labels identify the predicted class whenever it differs from the ground truth.	85
4.7	Visual confusion matrix with sample images along with the corresponding CAM.	86
5.1	Two samples of thickness measurement of metacarpal head (MH) cartilage by US. Yellow arrow: chondrosynovial interface. Green arrow: osteochondral interface.	90

List of Figures

5.2	Proposed framework pipeline. A convolutional neural network is trained to predict the distance field from each of margin of the interfaces taken into account for thickness computation (lightblue box). A threshold operation is computed to extract the superficial margin of the chondrosynovial and the superficial margin of the osteochondral profiles from distance field. Cartilage thickness is then computed between corresponding points in the upper and lower interfaces, from the left-end (yellow arrow) to the right-end of the upper interface (red box). The obtained thickness measurement (purple points) is compared to that obtained by an expert clinician (red points). . . .	92
5.3	Ground-truth generation for distance fields. The raw image is annotated by drawing points along the interfaces of interest. (a) Cartilage interfaces delineation is obtained by cubic spline interpolation. The distance field is built by mounting a Gaussian distribution (length = r) in the line that is normal to the tangent line at each cartilage point (b). This is done for both the upper (c) and lower (d) interfaces, obtaining a double channel ground.	94
5.4	Post-processing procedure. The predicted distance fields for the upper (a) and lower (b) interfaces are thresholded and skeletonized (c). Cartilage thickness is calculated as the Euclidean distance between the corresponding points of the skeletonized interfaces (c).	96
5.5	Challenges in the dataset include US artifacts (b-e), and anatomical (a-c) and cartilage position (d) variability.	96
5.6	Visual samples of results obtained from the ablation-study. Next to each image, shallow (1st column), Unet (2nd column), short-proposed (3rd column), middle-proposed (4th column) and proposed (5th column) prediction results along with ground true images (GT, 6th column) are shown. First row and second row related to each image refer to upper and lower cartilage surface distance fields, respectively.	99
5.7	Visual examples of thickness measurement obtained by the architectures of our ablation study and by an expert clinician performing the measurement for the second time. The sea water points are the ones annotated by the clinician and used as ground true. Red points are the closer points to the sea water point, as estimated by each architecture of the ablation study and second clinician annotation. Images are zoom in to better appreciate results.	100

List of Tables

2.1	Contingency table.	10
2.2	Performance metrics.	13
2.3	Publicly available datasets used for fetal ultrasound deep-learning algorithms development. For performance metrics refer to Table 2.2. *: for the HC18 challenge dataset, only the total number of subjects (551) is reported. MVP: maximum vertical pocket.	15
2.5	Summary of deep-learning algorithms for fetal standard-plane detection (for performance metrics refer to Table 2.2). FASP = Fetal Abdomen Standard Plane, FFSP = Fetal Facial Standard Plane, FFASP = Fetal Face Axial Standard Plane, 4CH = Four Chamber View, FBSP = Fetal Brain Standard Plane, LVOT = Left Ventricular Outflow Tract, 3VV = Three-Vessel View, RVOT = Right Ventricular Outflow Tract, FFESP = Fetal Femur Standard Plane, TT = Trans-Thalamic Standard Plane, TC = Trans-Cerebellum Standard Plane, FLVSP = Fetal Lumbosacral Spine Standard Plane, TV = Trans-Ventricular Standard Plane. Performance metrics are expressed as the average value over the classes. *: Only the total number of subjects is reported. □: The dataset includes unhealthy subjects. . .	18
2.6	Summary of deep-learning algorithms for anatomical structure analysis (for performance metrics refer to Table 2.2). *: Only the total number of subjects is reported. □: The dataset includes unhealthy subjects. .	23
2.7	Summary of deep-learning algorithms for biometry parameter estimation (for performance metrics refer to Table 2.2). HC: Head circumference, BPD: Biparietal diameter, OFD: occipito-frontal diameter, AC: Abdominal circumference, FL: Femur length, TCD: trans-cerebellar diameter. *: Only the total number of subjects is reported. □: The dataset includes unhealthy subjects.	39
3.1	Metric comparison for the ablation study. The mean is reported for each metric, with standard deviation in brackets.	65
3.2	Mask-R ² CNN distance-field regression heads. Conv = convolution, UpSamp = upsampling, $d \times d$ in the Output dimension column = spatial size of the squared feature map in output from the ROI align layer. The number of channels is reported, too.	69

List of Tables

3.3	Ablation-study results. The mean value, with standard deviation in brackets, is reported for each metric. All metrics but the Dice similarity coefficient are reported in mm.	73
4.1	Number of training images and patients for each class.	81
4.2	Number of images and patients for each class in the test set	81
4.3	Data augmentation used during contrastive learning (strong augmentation) and classification (weak augmentation). Ranges for each transformation are reported along with the augmentation probability.	82
5.1	Cartilages' regression network architecture with Conv2D = convolution, MaxPooling = max pooling, Upconv = up convolution, Concat = concatenation	95
5.2	Demographic characteristics of the subjects included in the study.	97
5.3	Ablation study summary. Encoder-decoder structure, atrous convolutions and depth (number of layers) are the model characteristics taken into account.	101
5.4	Ablation study comparison in terms of mean absolute error between distance field architectures predictions and distance field ground truth images.	102
5.5	Ablation study & intra-clinician variability assessment in terms of absolute difference and difference. We report the mean for each metric, with standard deviation in brackets. Metrics are calculated between the cartilage thickness obtained by the architecture (or for the second time by the same clinician) and clinician manual annotation.	103

Chapter 1

Introduction

1.1 Problem statement and motivation

Medical imaging plays a critical role in healthcare serving as an essential resource for early diagnosis, staging and management of different diseases as well as treatment selection, planning, guidance and follow up [3]. Due to its real-time nature, ultrasound (US) imaging stands out among many imaging diagnostic options as a highly cost-effective modality that provides the clinician with an unrivaled and invaluable level of interaction. Other advantages include the absence of ionizing radiation, as well as US portability, accessibility and cost effectiveness. Thanks to these characteristics, US has been first introduced in the obstetrics and gynecology and after that in almost all the fields of the medicine (abdominal US, breast US, echocardiography, transrectal US, intravascular US and so on) [4].

Despite the well-recognized clinical utility, US also presents unique challenges such as presence of speckle and acoustic noise which hinder image interpretation, high dependence on diagnostic operator experience and high inter- and intra-observer variability across different institutes and manufacturers' US systems. To overcome these challenges, different automated image analysis methods have been proposed over the year with the aim of making US diagnosis and assessment more objective, accurate and intelligent.

By extrapolating hidden patterns from data and exploiting its excellent ability to make decision rules [5], Deep Learning (DL) has set itself to become one of the most important tools in areas such as image recognition and computer vision. It comes as no surprise that in recent years, after establishing its efficacy in a number of medical imaging tasks and applications, DL has been acknowledged to be the first real step towards the so-called Artificial Intelligence (AI) era [6]. Despite its huge potential, there is still room for improvement. First of all, the design of DL architectures specifically built to accomplish medical image analysis tasks along with expert knowledge update state-of-the-art (SOTA) performance results (as also stated in [7]). Second of all, having DL already outperformed humans in certain domains [8], DL algorithms are also expected to have clinical implications.

This thesis aims at moving the DL SOTA forward for already investigated US fields

such as gynecology, a well-known and studied domain of application; but also it aims at investigating the potentiality of DL approaches for fields such as rheumatology in which DL has not been fully exploited.

1.2 Introduction to deep learning for ultrasound image analysis

When dealing with medical US image analysis, applications of DL include biometric measurements, computer-aided diagnosis, image-guided intervention and therapy which are most often solved by means of classification, detection and segmentation tasks in a supervised manner, so by inferring a mathematical relationship between the inputs and labeled outputs.

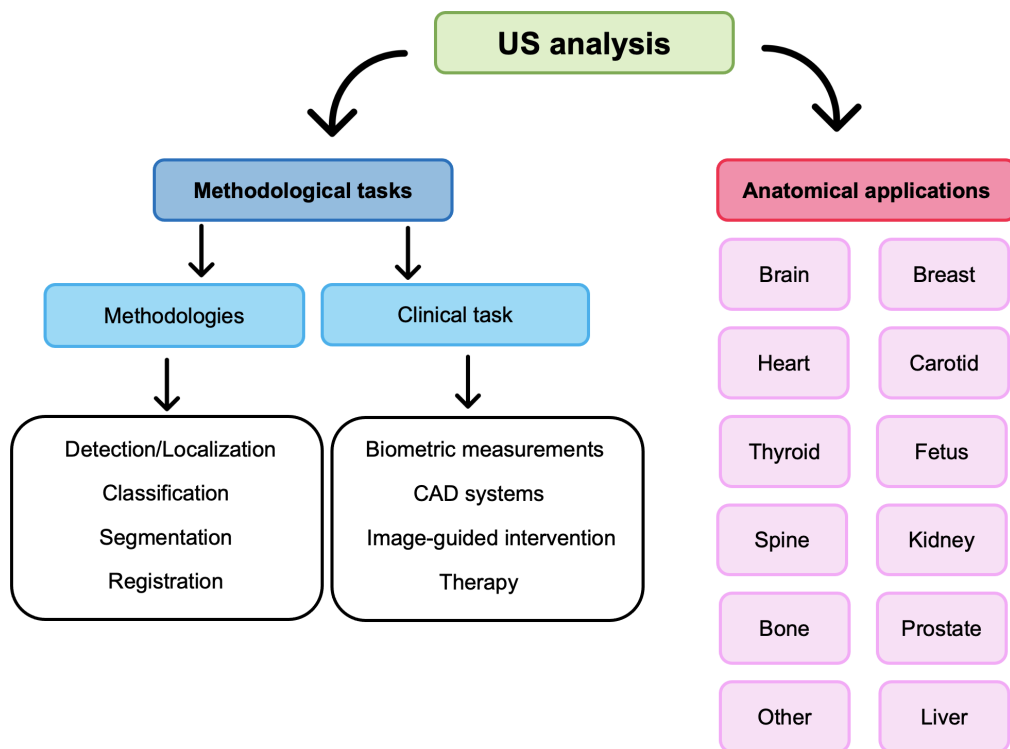


Figure 1.1: Illustration of medical US analysis surveyed by [1] .

Among the aforementioned tasks, classification includes a wide range of applications from determining the presence or absence of a target of interest (tumor, nodules, fetuses) up to identifying cardiac cycle phases (end-diastolic and end-systolic) for cardiac parameters estimation. Localization or detection of anatomical structures is another crucial step in US image analysis workflow: organs anatomies are often distinguished by clinicians by searching for specific anatomical signatures. Other detection tasks include the detection of lesions for therapy planning and intervention, malformations detection which influence prenatal management and the detection

of the specific fetal standard plane. As regards segmentation approaches, they are often used to quantitatively analyse clinical parameters such as volume and shape of organs, detect and classify lesions (e.g. breast, prostate, thyroid nodules etc) and extrapolate biometric measurements (ranging from infant to elderly).

These tasks are widely applied to different anatomical structures such as breasts, prostate, heart, liver, brain, carotid, thyroid, fetus, lymph node, kidney, spine, bone, muscle and nerve structure. Figure 1.1 shows the most investigated methodologies along with anatomical US applications according to one of the most updated reviews of the field [1].

Recently, US saw a rise in usage when it came to musculoskeletal assessments after undergoing considerable technological enhancements, as AI-based computer-aided detection and computer-aided diagnosis are fundamental to ensure less costly, more effective and better US imaging overall. [9]. The most DL investigated tasks are nerve identification and segmentation [10, 11, 11], myositis classification [12] and synovitis classification [13], yet the musculoskeletal domain remains an uncharted territory with respect to many applications such as cartilage thickness estimation in patients with Rheumatoid Arthritis (RA).

1.3 Aim of the thesis

The overall goal of this PhD thesis is to exploit DL algorithms for US image analysis. In particular, this work focuses on two specific fields: i) fetal US imaging, in which US is daily used in clinical practice and a huge research also including DL applications already exists; and ii) rheumatology, in which DL usage is quite innovative and thus there is still much to be done in relation to DL methodologies and applications.

In particular, starting from the analysis of the open technical challenges in gynecology field and the opportunities in an-explored ones, the research hypotheses that guide the PhD work can be summarized as follows:

- DL frameworks specifically designed to tackle different aspects of the problem at hand broaden the field's SOTA. This could be also done by taking advantage of publicly available datasets of the field.
- DL approaches that address problems with US technology that have not been solved yet incentivise research in the field and it could also benefit from a clinical point of view

The methodology developed to investigate the hypotheses will be described in this PhD dissertation. Sample applications will be proposed for experimentally testing the hypotheses.

1.4 Structure of the thesis

This dissertation is organized as follows:

Chapter 2 focuses on the deep understanding of fetal ultrasound DL methodologies. This chapter ends with a critical summary of what is currently missing in the field.

In Chapter 3, proposed DL methodologies specifically design to automatically delineate head circumference (HC) from fetal US images are presented. Each strategy is developed using a available dataset of the field, allowing the direct comparisons with SOTA algorithms.

In Chapter 4, a new semi-supervised approach for fetal standard plane detection is proposed by taking advantage of the only available dataset of the field.

Chapter 5 describes a new DL method for the cartilage thickness estimation from US rheumatological images, paving the way for future research in the field

In Chapter 6, the conclusion, scientific and clinical implications, and future perspective of this PhD work are reported and discussed. The graphical abstract of this PhD thesis is shown in Figure 1.2.

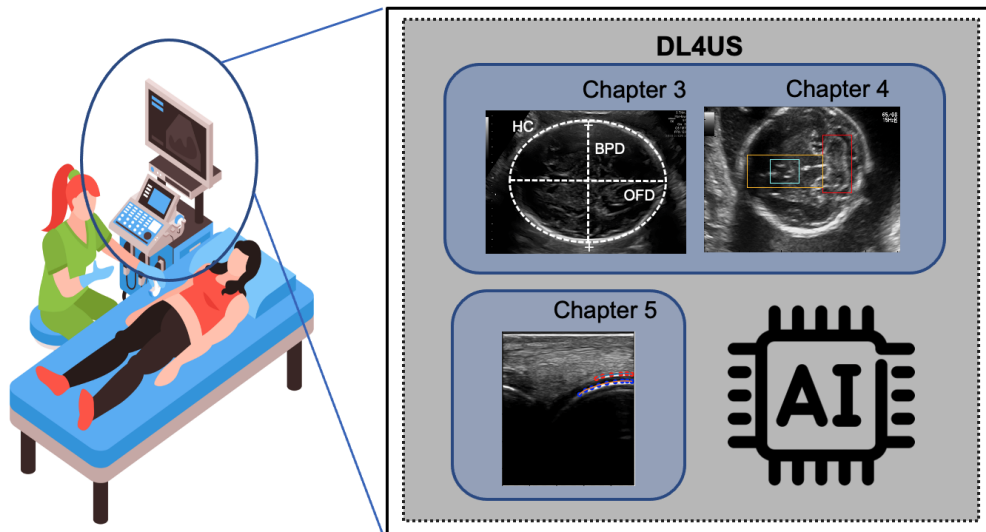


Figure 1.2: Graphical abstract of this PhD thesis. The aim is to develop automatic algorithms which may help clinicians having more reliable and robust US image analysis. In particular, Chapter 3 and 4 present fetal US DL algorithms to delineate HC and perform scan plane classification, respectively. In Chapter 5 a DL algorithm for cartilage thickness estimation is described.

1.5 Scientific publications

Part of the methodologies presented and tested in this PhD dissertation were presented in the following peer-reviewed publications.

1.5.1 Journal publications

Fiorentino, M. C., Villani, F.P, Di Cosmo, M., Frontoni, E. and Moccia, S. (2023). "A Review on Deep-Learning Algorithms for Fetal Ultrasound-Image Analysis." *Medical Image Analysis* 83, 102629 (2023)

Fiorentino, M. C., Cipolletta, E., Filippucci, E., Grassi, W., Frontoni, E., & Moccia, S. (2022). A deep-learning framework for metacarpal-head cartilage-thickness estimation in ultrasound rheumatological images. *Computers in Biology and Medicine*, 141, 105117.

Di Cosmo, M., **Fiorentino, M. C.**, Villani, F. P., Frontoni, E., Smerilli, G., Filippucci, E., & Moccia, S. (2022). A deep learning approach to median nerve evaluation in ultrasound images of carpal tunnel inlet. *Medical & Biological Engineering & Computing*, 60(11), 3255-3264.

Smerilli, G., Cipolletta, E., Sartini, G., Moscioni, E., Di Cosmo, M., **Fiorentino, M. C.**, ... & Filippucci, E. (2022). Development of a convolutional neural network for the identification and the measurement of the median nerve on ultrasound images acquired at carpal tunnel level. *Arthritis Research & Therapy*, 24(1), 1-8.

Moccia, S., **Fiorentino, M. C.**, & Frontoni, E. (2021). Mask-R² CNN: a distance-field regression version of Mask-RCNN for fetal-head delineation in ultrasound images. *International Journal of Computer Assisted Radiology and Surgery*, 16(10), 1711-1718.

Fiorentino, M. C., Moccia, S., Capparuccini, M., Giamberini, S., & Frontoni, E. (2021). A regression framework to head-circumference delineation from US fetal images. *Computer Methods and Programs in Biomedicine*, 198, 105771.

Cipolletta, E., **Fiorentino, M. C.**, Moccia, S., Guidotti, I., Grassi, W., Filippucci, E., & Frontoni, E. (2021). Artificial intelligence for ultrasound informative image selection of metacarpal head cartilage. A pilot study. *Frontiers in Medicine*, 8, 589197.

Conference proceedings

Migliorelli, G. & **Fiorentino, M.C.**, Villani, F.P., Di Cosmo, M., Frontoni, E., and Moccia, S. (2023) "On the use of contrastive learning for standard-plane classification in fetal ultrasound imaging" and it is currently under review for The 14th International Conference on Information Processing in Computer-Assisted Interventions (IPCAI)

Chapter 1 Introduction

Di Cosmo, M., **Fiorentino, M. C.**, Villani, F. P., Sartini, G., Smerilli, G., Filippucci, E., ... & Moccia, S. (2021, November). Learning-Based Median Nerve Segmentation From Ultrasound Images For Carpal Tunnel Syndrome Evaluation. In 2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC) (pp. 3025-3028). IEEE.

This thesis had also technology transfer implications. The author had the opportunity to join Esaote R&D and Vicomtech V6 groups. Esaote¹ is one of the world's leading producers of medical diagnostic systems and internationally acknowledged to be the world leader in dedicated MRI and US. The Esaote Group is also one of the main players in the sector of Information Technology for healthcare. Vicomtech² is a technological centre set up as a private non-profit Foundation. It responds to Applied Research, Development and Innovation in Information Technology, especially the convergence of Computer Graphics and Computer Vision (Visual Computing), Data Analytics and Intelligence, Interactive Digital Media and Language Technologies, in businesses and institutions in biomedical domain. During these months, the author had the possibility to focus on implementing and validate AI-based solutions for US and MRI.

¹<https://www.esaote.com/>

²<https://www.vicomtech.org/>

Chapter 2

Literature review on fetal-ultrasound analysis

Deep-learning (DL) algorithms are becoming the standard for processing ultrasound (US) fetal images. Despite the large number of survey papers already present in this field, most of them are focusing on a broader area of medical image analysis or not covering all fetal US DL applications

This chapter surveys the most recent work in the field, with a total of 153 research papers published after 2017 with the aim of increasing DL fetal US applications understanding. Each paper is analyzed and commented from both the methodology and application perspective. Papers are categorized in (i) fetal standard-plane detection, (ii) anatomical-structure analysis and (iii) biometry parameter estimation. For each category, main limitations and open issues are presented. Summary tables are included to facilitate the comparison among the different approaches. Publicly-available datasets and performance metrics commonly used to assess algorithm performance are summarized, too. This chapter ends with a critical summary of the current state of the art on DL algorithms for fetal US image analysis and a discussion on current challenges that have to be tackled by researchers working in the field to translate the research methodology into the actual clinical practice.

A journal paper on this work has been published as [14]: Fiorentino, M.C., Villani, F.P., Di Cosmo, M., Frontoni, E. and Moccia, S. (2023). "A Review on Deep-Learning Algorithms for Fetal Ultrasound-Image Analysis." *Medical Image Analysis* 83, 102629 (2023)

2.1 Introduction

Ultrasound (US) imaging is an imaging modality widely used for the diagnosis, screening and treatment of a large number of disease, due to its portability, low cost and non-invasive nature [15]. In the years, US imaging has turned out to be the preferred checkup method during pregnancy [16]. It is commonly used to evaluate fetus's growth and development, as well as to monitor pregnancy and assess clinical suspicion[17]. Traditionally, 2-dimensional (2D) US, which produces 2D images that dynamically display fetus's body and internal organs in real-time, is used during

pregnancy. More recently, 3-dimensional (3D) US has been made available to support clinicians to better understand fetal spatial anatomic relationships as well as to closely examine suspected fetal anomalies [18].

From the clinician's perspective, analyzing fetal US images may be challenging due to the presence of artifacts, such as acoustic shadows, speckle noise, motion blurring and missing boundaries, which are produced as the result of the complex interaction between US waves and mother and fetal biological tissues [19]. Further challenges include fast fetus movements, occluded anatomical structures (e.g. due to fetus positioning) and high variability associated with different gestational weeks.

This Chapter describes a review that surveys the most recent work in the field of fetal US image analysis with DL. This review has the aim to be a valuable and compact source of information for young researchers, and a reference overview document for those already working in the field.

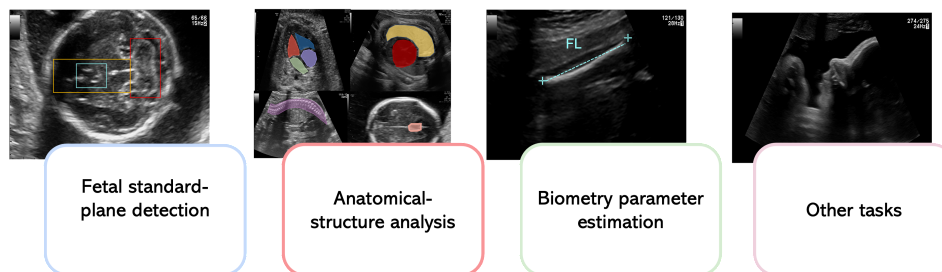


Figure 2.1: Summary of the tasks surveyed.

During the last decades, DL, and in particular convolutional neural networks (CNNs), have undergone, indeed, an increasing role in fetal US image analysis to offer decision support to clinicians, and today an extensive literature exists. Survey papers in the field have been published in the last years, even if a number of them focuses on the broader area of medical-image analysis [1] or surveys also DL algorithms for US image analysis outside the fetal field ([3, 20, 21, 22, 15]. Survey papers specifically dealing with US fetal images include: [23, 24] and [25], where segmentation and classification algorithms are covered; [26] and [27] that survey methods for fetal cardiology images; [28] and [29], that briefly summarize DL methods for fetal abnormality detection; and [30] and [31], that analyze research papers from a clinical perspective.

This review starts describing commonly used metrics used for algorithm performance assessment, as well as publicly-available datasets in the field (Sec. 2.2.2). As shown in Fig. 2.1, research papers ranging from fetal standard-plane detection (Sec. 2.3) to anatomical structure analysis (Sec. 2.4) and fetal-biometry estimation (Sec. 2.5) are surveyed. These sections mirror the steps of what is currently done in

clinical practice to evaluate fetus well-being. A miscellaneous section (Sec. 2.6) is also included, collecting papers on emerging tasks, from less common fetus evaluation applications to probe movement control. For each section, methods are described highlighting pros and cons. Limitations and open issues are further discussed. Summary tables are included to report information on training and testing datasets, as well as gestational age (GA) and achieved performance. A discussion on future directions and open challenges in the field of fetal US analysis with DL concludes this review (Sec. 3.5).

A journal paper on this work has been published as [14]: Fiorentino, Maria Chiara, et al. "A Review on Deep-Learning Algorithms for Fetal Ultrasound-Image Analysis."

2.1.1 Survey strategy

The survey strategy started with the following research questions:

- Which are the most investigated tasks addressed using DL in the field of fetal US image analysis?
- Which are the main challenges in regard to fetal examination that are currently tackled by using DL?
- Are the commonly-used datasets sufficient enough for robust DL algorithm development and testing?
- Which are the open issues that still have to be addressed by DL in the field?

With these questions in mind, we outlined a set of keywords for our survey, including: *classification, detection, segmentation, fetal, ultrasound, deep learning* combined together with terms related to fetal examination and organs. The research databases were IEEEXplore, Scopus, SpringerLink, Sciencedirect and Pubmed. For each resulting paper, an extensive review of its reference list was performed. To better focus on the most recent and interesting trends and to not overlap with previous review work, these following criteria were considered:

- Papers proposing DL methodologies to process fetal US images
- Indexed papers (Scopus and Web of Science)
- Papers published after 2017
- In regard to journal papers, we only considered papers published in journals ranked (at least) in the 50th percentile, according to CiteScore rank 2020

A summary of the characteristics of the papers analyzed in this review is shown in Fig. 2.2. The survey resulted in 153 papers.

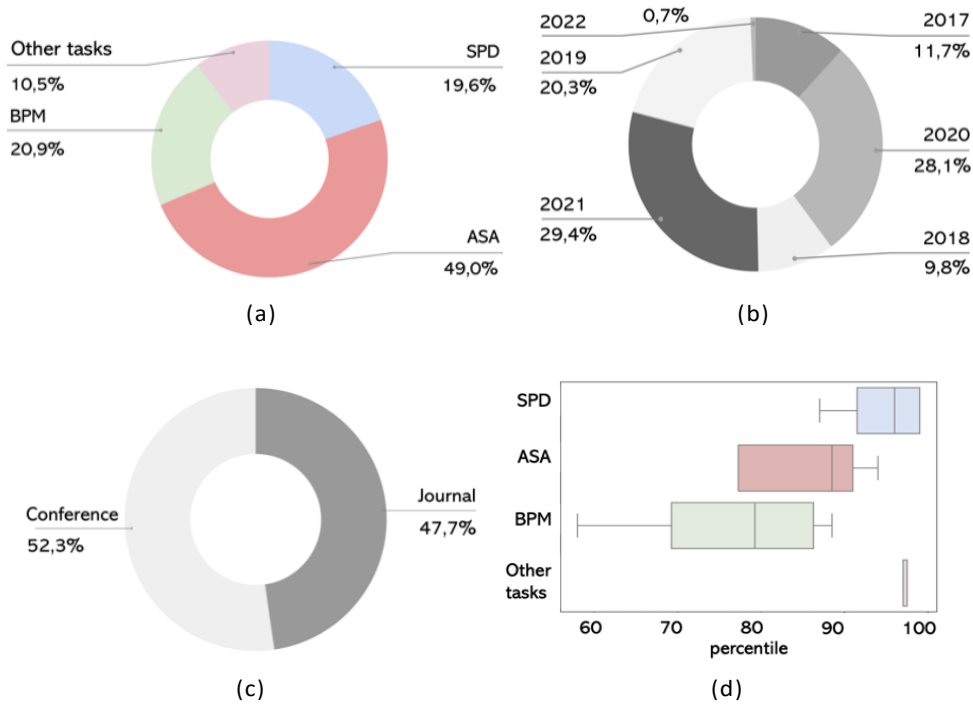


Figure 2.2: Summary of the papers included in this work considering: (a) addressed task, (b) year of publication, (c) contribution (journal/ conference), (d) percentile according to CiteScore rank 2020 (only for journal papers), SPD = standard-plane detection, ASA = anatomical structure analysis, BPM = biometry parameter estimation.

Table 2.1: Contingency table.

		Gold Standard	
		Yes	No
Algorithm result	Yes	TP	FP
	No	FN	TN

2.2 Metrics and datasets

2.2.1 Performance metrics

DL algorithms for fetal US image analysis are evaluated using different performance metrics according to the addressed task. Main tasks include classification, which is particularly used for standard plane detection (Sec. 2.3) and anatomical structure analysis (Sec. 2.4), especially when providing diagnostic support. Besides classification, segmentation and localization/detection are often investigated. The former finds its main application in anatomical structure analysis (Sec. 2.4) and biometry estimation (Sec. 2.5), as a prior for performing computer-assisted diagnosis (CAD) and/or measuring biometry parameters. Localization and detection are commonly used to localize structures of interest in the image, which finds application in standard

plane detection, anatomical structure analysis and biometry estimation.

This section presents the most used quantitative metrics for assessing DL algorithm performance, with the goal to promote fair comparisons. Publicly-available datasets for algorithm development and testing are covered as well. To limit the overlap with the existing literature, we avoid to report on loss functions, as standard loss functions are commonly used in the field of DL for fetal US image analysis. Comprehensive surveys in the field may be found in [32] and [33].

For classification tasks, performance is assessed by means of the contingency table (Table 2.1), with True Positives (TP), True Negatives (TN), False Negatives (FN) and False Positives (FP). Popular classification metrics computed from the contingency table are: (i) Accuracy (Acc) = number of correct predictions ($TP + TN$) divided by the total number of predictions (N); (ii) Recall (Rec) = fraction of actual positives which are correctly identified; (iii) Specificity ($Spec$) = fraction of actual negatives which are correctly identified; (iv) Precision ($Prec$) = proportion of positives which are identified. The F1-score ($F1$) metric, which is equal to the harmonic mean of $Prec$ and Rec , is often used, too. These metrics can be computed at patch, image or patient level.

Other popular metrics are top-1 and top-3 error rates. For top-1 error rate, only the top class (i.e., the one with the highest output probability) is compared with the target label. In the case of top-3 error rate, the target label is compared with the top three predictions (i.e., the three with the highest probability). In both cases, the top score is computed as the number of times the predicted label matched the target label, divided by N , total number of predictions.

The Receiver Operating Characteristic (ROC) curve, which shows the performance of a binary classifier as a function of its cut-off threshold, is often reported. The area (AUC) under the ROC curve is used as metric and is interpreted as the probability that the DL model ranks a random positive example more than a random negative example. The higher the AUC (close to 1), the better the model performance.

Even if Acc is widely used, it is not representative when dealing with imbalanced datasets. Therefore, the $F1$ is adopted for imbalanced datasets and multi-class problems. Using the AUC should also be avoided in case of imbalanced datasets, where $Prec \square Rec$ curve should be used instead [34].

As regards segmentation tasks, model performance is commonly evaluated by means of the Dice Similarity Coefficient (DSC) and Intersection over Union (IoU). DSC is equivalent to $F1$ and can be defined as:

$$DSC = \frac{2|X \cap Z|}{|X| + |Z|} = \frac{2TP}{FP + FN + 2TP} = F1 \quad (2.1)$$

where X and Z are the predicted and the labeled masks, respectively. The IoU is defined as:

$$IoU = \frac{|X \cap Z|}{|X \cup Z|} \quad (2.2)$$

Chapter 2 Literature review on fetal-ultrasound analysis

The Hausdorff distance (HD), which measures how far the labeled segmentation mask is from the predicted segmentation, can be used, too:

$$HD(X, Z) = \max(h(X, Z), h(Z, X)) \quad (2.3)$$

where

$$h(X, Z) = \max_{x \in X} \max_{z \in Z} \|x - z\| \quad (2.4)$$

DSC and IoU are positively correlated, however IoU tends to penalize single instances of bad segmentation more than DSC , which tends to measure something closer to average performance. Besides DSC and IoU , HD is especially recommended for segmentation tasks with complex boundaries and small thin segments [35], even if it does not consider progress of curves.

For localization tasks, performance is assessed by means of the Average Precision (AP), which is the $Prec$ averaged across all Rec values between 0 and 1. AP may be seen as the area under the precision-recall curve. The $Prec$ and Rec are commonly computed at various thresholds of IoU between the predicted and labeled bounding box. For detection tasks, where multiple structures have to be detected in the image, mean AP (mAP) is used:

$$mAP = \frac{\sum_{i=1}^K AP_i}{K} \quad (2.5)$$

where K is the number of classes.

For regression tasks, where a numerical value has to be predicted (e.g., for biometry estimation), popular metrics are:

- Mean squared error (MSE), which computes the average squared error between the predicted (y_i) and actual values ($\hat{y}_i$):

$$MSE = \frac{1}{N} \sum_{i=1}^N (y_i - \hat{y}_i)^2 \quad (2.6)$$

Due to its differentiable nature, it is often used as loss function during training.

- Mean absolute error (MAE), sometimes referred to as mean absolute deviation, defined as the average of the absolute distance between y_i and $\hat{y}_i$:

$$MAE = \frac{1}{N} \sum_{i=1}^N |y_i - \hat{y}_i| \quad (2.7)$$

- Difference error (DF):

$$DF = \frac{1}{N} \sum_{i=1}^N (y_i - \hat{y}_i) \quad (2.8)$$

Table 2.2: Performance metrics.

Index	Description
Accuracy (<i>Acc</i>)	$\frac{TP + TN}{N}$
Recall (<i>Rec</i>)	$\frac{TP}{TP + FN}$
Specificity (<i>Spec</i>)	$\frac{TN}{TN + FP}$
Precision (<i>Prec</i>)	$\frac{TP}{TP + FP}$
top-1 error rate / top-3 error rate	Top-1 and top-3 errors
AUC	Area Under the Receiver Operating Characteristic curve
AUC by Judd	AUC variant from [36]
Dice similarity coefficient (<i>DSC</i>)	Eq. 2.1
Intersection Over Union (<i>IoU</i>)	Eq. 2.2
Hausdorff distance (<i>HD</i>)	Eq. 2.3
mean Average precision (<i>mAP</i>)	Eq. 2.5
Mean squared error (<i>MSE</i>)	Eq. 2.6
Mean absolute error (<i>MAE</i>)	Eq. 2.7
Difference (<i>DF</i>)	Eq. 2.9
Euclidean distance (<i>ED</i>)	Eq. 2.10
Kullback-Leiber divergence (D_{KL})	Eq. 2.11
Normalized Scanpath Saliency (<i>NSS</i>)	Eq. 2.12

- Absolute difference error (*ADF*):

$$DF = \frac{1}{N} \sum_{i=1}^N (y_i - \hat{y}_i) \quad (2.9)$$

- Euclidean distance (*ED*):

$$ED = \frac{1}{N} \sum_{i=1}^N \sqrt{(y_i - \hat{y}_i)^2} \quad (2.10)$$

MAE is reported along with *MSE* because it is more robust to outliers. Hence, by relying on the absolute error, *MAE* avoids errors to cancel each other out as when calculating the *MSE*. Furthermore, *MAE* does not penalize larger errors more than smaller ones. The advantage of *DF* and *ED* over *MSE* is that, as for *MSE*, they are measured in the same units as the variable to be regressed.

Recently, researchers are focusing on visual saliency prediction (i.e., predicting

human eye fixations on images in the form of saliency maps). Models for saliency prediction can be evaluated using a variety of performance metrics, as described in [37], which provides an in-depth overview of their advantages and disadvantages. Among them, in the field of US fetal image analysis the Kullback-Leiber divergence (D_{KL}) is often used, which is defined as:

$$D_{KL}(S || F^D) = \sum_{i=1}^T F_i^D \log(\epsilon + \frac{F_i^D}{\epsilon + S_i}) \quad (2.11)$$

where S and F^D are the (predicted) saliency map and (ground truth) continuous fixation map distributions, respectively, ϵ is a regularization constant and T is the total number of fixated pixels.

The Normalized Scanpath Saliency (NSS) [38] is used, too. The NSS is computed as:

$$NSS(S, F^B) = \frac{1}{T} \sum_{i=1}^T \bar{S} F_i^B \quad (2.12)$$

where F^B is the ground truth (binary) fixation location map and

$$\bar{S} = \frac{S - \mu(S)}{\sigma(S)} \quad (2.13)$$

An AUC variant from [36] is also used. Here the AUC is built considering TP and FP rates defined as follows. For a given threshold, the TP rate is the ratio of true positives to the total number of fixations, where true positives are saliency map values above threshold at fixated pixels. The FP rate is the ratio of false positives to the total number of saliency map pixels at a given threshold, where false positives are saliency map values above threshold at unfixed pixels.

Table 2.2 summarizes the performance metrics described in this section.

2.2.2 Publicly-available datasets

Collecting and sharing high-quality annotated fetal US datasets is not trivial. Labeling large datasets can take a significant amount of time, which may vary according to the task (e.g., pixel-level labeling for segmentation is way more time-consuming than image-level labeling for classification). Data privacy and protection concerns further constitute a barrier to data sharing. To attenuate these issues, international scientific organizations are working to collect and share publicly available databases to encourage algorithm development and fair comparison among algorithms. In the framework of the IEEE International Symposium on Biomedical Imaging (ISBI) and the International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI), three annotated datasets have been released in the form of a Grand Challenge¹.

¹<https://grand-challenge.org/challenges/>

2.2 Metrics and datasets

Table 2.3: Publicly available datasets used for fetal ultrasound deep-learning algorithms development. For performance metrics refer to Table 2.2. *: for the HC18 challenge dataset, only the total number of subjects (551) is reported. MVP: maximum vertical pocket.

Name	Task(s)	Training set size (images/ patients)		Testing set size (images/ patients)		Annotators	Performance metrics	Gestational age
Challenge US (2012)	Head and femur segmentation & measurement	-/-		284 2D/-		2 and 3, respectively	<i>Prec</i> [%], <i>Acc</i> [%], <i>DSC</i> [%] and distance-based metrics	21, 28 and 33 weeks
HC18 challenge dataset (2018)	Head-circumference estimation	999 2D/ *		335 2D/ *		1	<i>MAE</i> [mm], <i>DF</i> [mm], <i>DSC</i> [%], <i>HD</i> [mm]	12 to 35 weeks
A-AFMA ultrasound challenge dataset (2021)	1) Amniotic fluid and maternal bladder detection 2) MVP detection	-		-		-	<i>mAP</i> [0-1]	-
[39]	Standard plane detection of abdomen, brain, heart, femur, maternal cervix and other	7129 896	2D/	5271 896	2D/	1	top-1 error rate [%], top-3 error rate [%], <i>Acc</i> [%]	18 to 40 weeks

Chapter 2 Literature review on fetal-ultrasound analysis

The first challenge in time was the Challenge US: Biometric Measurements from Fetal Ultrasound Images [40], held in conjunction and with the support of ISBI in 2012. The goal of the challenge was to automatically segment fetal abdomen, head, femur and whole fetus for measuring standard biometric parameters, but none attempted the abdomen and whole fetus sub-challenges. Despite the recognized value of the dataset, its size (270 images) does not allow researchers to develop generalizable DL algorithms.

In 2018, with the release of the HC18 challenge dataset² [41], the potential of DL for biometry parameter estimation has been unlocked. The challenge was to develop algorithms to automatically measure fetal head circumference (HC), with 999 and 335 2D US images for training and testing, respectively. Images were annotated by an experienced sonographer. For a more in depth description refer to 3.2.2.

During ISBI 2021, the A-AFMA ultrasound challenge³ was organized. The goals were to: 1) detect amniotic fluid and maternal bladder, 2) identify the appropriate landmarks for maximum vertical pocket (MVP) measurement, as to assess amniotic fluid volume.

Recently, a large dataset⁴ of routinely acquired maternal-fetal screening US images was made publicly available in [39]. It consists of 7129 2D training images from 896 patients, categorized into 6 classes: abdomen, brain, femur, thorax, maternal cervix and other. Images were manually labeled by an expert fetal clinician. A test set, which consists of 5271 2D images from 896 patients, was provided. The goal was to boost the research in the field of fetal standard-plane detection.

The publicly-available datasets for the development of DL algorithms for fetal US are summarized in Table 2.3

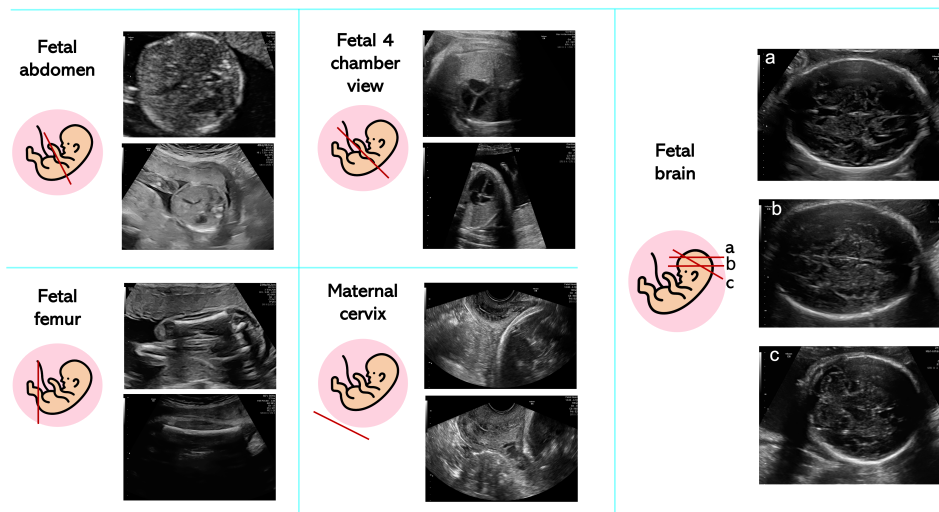


Figure 2.3: Visual samples of the most common fetal standard planes.

²<https://saras-mesad.grand-challenge.org/>

³<https://a-afma.grand-challenge.org/>

⁴<https://zenodo.org/record/3904280>

2.3 Fetal standard-plane detection

According to the International Society of Ultrasound in Obstetrics and Gynecology (ISUOG) guidelines [71], the use of standardized planes of acquisition improves the reproducibility of (i) fetal biometry assessment and (ii) fetus evaluation.

The planes that are typically acquired to extrapolate biometric measurements are fetal abdomen (FASP), brain (FBSP) and femur (FFESP) standard planes. FBSP involves trans-ventricular (TV) and trans-thalamic (TT) standard planes. Fetus evaluation further requires the acquisition of maternal cervix, fetal heart (including 4 chamber view (4CH), left ventricular outflow tract (LVOT), right ventricular outflow tract (RVOT), three-vessel trachea (3VT), three-vessel view (3VV)), fetal trans-cerebellum (TC) standard plane, fetal facial standard plane (FFSP) and lumbosacral spine plane (FLVSP). FFSP includes axial (FFASP), coronal and sagittal planes.

Visual samples of the most common standard planes are shown in Fig. 2.3.

In the clinical practice, the acquisition of a standard plane is performed manually by clinicians, which move the US probe across the mother's belly until specific anatomical landmarks are visible in the image. Clinical expertise is required to face the high intra-class variability of US standard planes, which is due to, among the others, different gestational weeks, equipment vendors and US-probe angle [52]. Moreover, the anatomical structures that characterize a specific plane could be common to other planes. While approaches in the last decade focused on handcrafted feature-based machine learning [72], today DL algorithms are proving to be a valuable tool to tackle these challenges. Table 2.5 summarizes the DL algorithms for standard-plane detection that are surveyed in this section. An overview of the most common strategies adopted in the literature is shown in Fig. 2.4.

One of the first DL algorithms for scan plane detection is proposed by [42], with the goal of detecting FASP. The algorithm consists of a first CNN that localizes fetal abdomen and a second one that detects the presence in the localized abdomen of stomach bubble and umbilical vein. Despite the promising results, two separate CNNs are needed for localization and classification, respectively, inevitably increasing training and deployment time.

A number of algorithms tackles the problem of fetal standard plane detection as a mere classification problem of 2D still images, following the strategy shown in Fig. 2.4a. In [43], FFSP is detected by fine tuning a shallow classification CNN pre-trained on ImageNet⁵. A shallow classification CNN is also used in [44] to automatically identify FBSPs. In [39], state-of-the-art CNNs are compared to classify 6 planes (FASP, TT, TC, TV, FFESP, and thorax standard planes). The best performing CNN results to be DenseNet-169. The authors made the dataset available to foster research in the field (Sec. 2.2.2, Table 2.3). A dense network is also used in [45] to detect 4CH, FASP, FBSP, FFSPs. Similarly, the work in [46] proposes an automatic fetal standard plane classification of 4CH, FASP, FBSP, FFASP, and

⁵<https://www.image-net.org/>

Chapter 2 Literature review on fetal-ultrasound analysis

Table 2.5: Summary of deep-learning algorithms for fetal standard-plane detection (for performance metrics refer to Table 2.2). FASP = Fetal Abdomen Standard Plane, FFSP = Fetal Facial Standard Plane, FFASP = Fetal Face Axial Standard Plane, 4CH = Four Chamber View, FBSP = Fetal Brain Standard Plane, LVOT = Left Ventricular Outflow Tract, 3VV = Three-Vessel View, RVOT = Right Ventricular Outflow Tract, FFESP = Fetal Femur Standard Plane, TT = Trans-Thalamic Standard Plane, TC = Trans-Cerebellum Standard Plane, FLVSP = Fetal Lumbosacral Spine Standard Plane, TV = Trans-Ventricular Standard Plane. Performance metrics are expressed as the average value over the classes. *: Only the total number of subjects is reported. □: The dataset includes unhealthy subjects.

Paper	Plane	Training set size	Test set size	Annotators	Performance metrics	Gestational age
[42]	FASP	8072 2D (492 videos) (-)	2606 2D (219 videos) (66 subjects)	3	AUC=0.99, Acc=0.98, Rec=0.96, Spec=0.97	16 to 40 weeks
[43]	FFSP	4849 2D□	2418 2D□	Few	AUC=0.99, Acc=0.96, Prec=0.96, Rec=0.97, F1=0.97	20 to 36 weeks
[44]	FBSPs	15314 2D (155 sub-jects*)	3828 2D (155 subjects*)	Few	Acc=0.93, Prec=0.93, Rec=0.92, F1=0.93	-
[39]	Multiple	7129 2D (896 subjects*)	5271 2D (896 subjects*)	1	6.2% top-1 error, 0.27% top-3 error, Acc = 0.94	18 to 40 weeks
[45]	4CH, FASP, FBSP, FF-SPs	17036 2D	5678 2D	1	Prec = 0.98, Rec = 0.98, F1 = 0.98	14 to 28 weeks
[46]	4CH, FASP, FBSP, FFASP, coronal FFSP	17840 2D	4455 2D	-	Acc = 0.99, Rec = 0.96, Spec = 0.99, F1 = 0.95	-
[47]	4CH, LVOT, 3VV	10000 2D (10 subjects)	2178 2D (2 subjects)	1	error rate = 0.23	20 to 35 weeks
[48]	FBSP	6498 2D	2249 2D	Few	Acc= 0.81, AUC = 0.86, F1= 0.80	-
[49]	Multiple	6974(920 sub-jects*)	2768 (920 subjects*)	1	Acc= 0.90, mAP= 0.94, Prec= 0.92, Rec= 0.94, F1= 0.93	18 to 28 weeks
[50]	4CH, FASP, LVOT, RVOT, FFESP, Lips	12000 2D	5500 2D	Few	F1=0.77, Rec=0.77, Prec= 0.78	-
[51]	FASP, FFASP, 4CH	37376 2D (900 sub-jects)	13248 2D (331 sub-jects)	1	Acc=0.87, Prec=0.71, Rec=0.64, F1=0.64	18 to 40 weeks
[52]	FASP, TT, TC, FLVSP	68296 2D (1199 videos)	16740 2D (244 videos)	Few	Acc = 0.85, Prec = 0.85, Rec = 0.85, F1 = 0.85	-

2.3 Fetal standard-plane detection

Paper	Plane	Training set size	Test set size	Annotators	Performance metrics	Gestational age
[53]	Multiple	122233 2D	38243 2D	Few	$Acc = 0.98$, $Rec = 0.93$, $Prec = 0.93$, $F1 = 0.93$	18 to 22 weeks
[54]	FASP	1292 2D (25 subjects)	324 2D (8 subjects)	Few	$Prec = 0.96$, $Rec = 0.96$, $F1 = 0.96$	-
[55]	FASP, FBSP, FFESP	224 videos	56 videos	1	$Prec = 0.98$, $Rec = 0.85$, $F1 = 0.87$	-
[56]	Multiple	1504 2D	752 2D	-	$Prec = 0.75$, $Rec = 0.73$, $F1 = 0.74$	mid-trimester
[57]	4CH	5626 2D	1406 2D	Few	$mAP = 0.81$	14 to 28 weeks
[58]	FBSP	4800 2D	1153 2D	1	$AP=0.79$, $Rec=0.85$, $Prec=0.87$	14 to 28 weeks
[59]	FBSP	1451 2D	320 2D	Few	$mAP = 0.93$	14 to 28 weeks
[60]	FASP, FBSP, 4CH	2460 2D	820 2D	2	$mAP=0.95$, $Acc=0.95$, $Prec=0.95$, $Rec=0.93$	20 to 34 weeks
[61]	Multiple	140827 2D and 2438 videos (2694 subjects*)	109165 2D and 200 videos (2694 subjects*)	45	$Prec = 0.77$, $Rec = 0.90$, $F1 = 0.80$ and $IoU = 0.62$	18 to 22 weeks
[62]	TV	15870 2D	3968 2D	Few	$Acc=0.95$	mid-gestation
[63]	TT, TC	34586 2D from 441 videos (147 subjects)	60 videos (20 subjects)	2	$mAP=0.87$	-
[64]	Multiple	140827 2D and 2438 videos (2694 subjects*)	109165 2D and 200 videos (2694 subjects*)	45	$Prec=0.89$, $Rec=0.90$, $F1=0.89$	10 to 22 weeks
[65]	Multiple	22757 2D (2000 subjects*)	5737 2D (2000 subjects*)	-	$Acc=0.70$	-
[66]	FBSPs	330 3D (330 subjects)	100 3D (100 subjects)	1	$DF=3.03$ mm, $\vartheta=9.36^\circ$	19 to 31 weeks
[67]	Multiple	1281 3D	354 3D	4	$DF=2.31$ mm, $\vartheta=10.36^\circ$	-
[68]	Uterine standard planes	539 3D (476 subjects*)	144 3D (476 subjects*)	4	$DF=1.82$ mm, $\vartheta=7.20^\circ$	19 to 31 weeks
[69]	TV, TC	50 3D (50 subjects)	22 3D (22 subjects)	1	mean plane centre difference = 3.44 mm, rotation angle = 11.05°	-
[70]	Middle sagittal plane	112 3D	28 3D	-	$ED = 0.05$	11 to 13 weeks

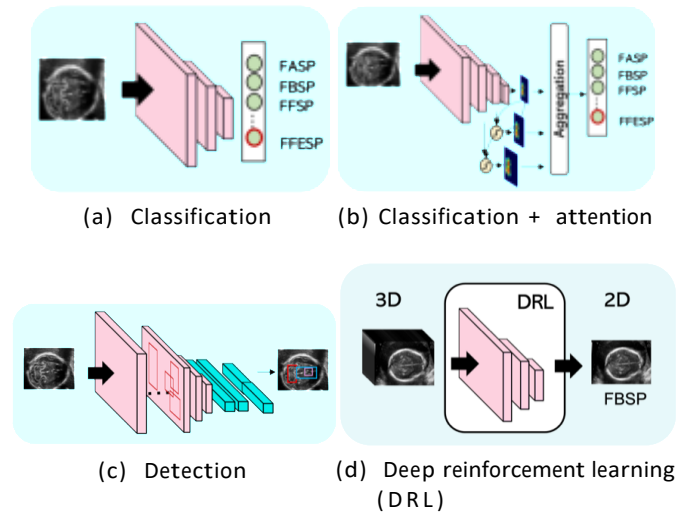


Figure 2.4: Most common deep-learning strategies to fetal standard-plane detection. FASP = fetal abdomen standard plane, FBSP = fetal brain standard plane, FFSP = fetal facial standard plane, FFESP = fetal femur standard plane.

coronal FFSP, based on DenseNet. The network is trained in conjunction with a placenta transferring dataset in order to discover and learn potential relationships between the datasets and to possibly avoid overfitting. In [47], the detection of heart fetal standard planes (4CH, LVOT, 3VV and not heart) is addressed with a Fully Convolutional Network (FCN), that aims at segmenting the center of fetal heart and classifying the cardiac views in a single step. A generative adversarial network (GAN) is used in [48] to improve FBSP classification by means of ResNet. In [49], a multi-label learning approach is proposed to simultaneously identify multiple standard planes and corresponding anatomical structures. The pipeline is threefold: first a word embedding is used to identify class dependency between standard planes and anatomical structures; then, a graph convolutional network is used to refine the dependency among classes; finally, contrastive learning is used to further increase the discriminative ability. Cross-device classification of 6 anatomical standard planes (4CH, FASP, LVOT, RVOT, FFESP, Lips) is performed in [50]. An improved feature alignment is used to extract discriminative and domain-invariant features across domains.

A couple of research papers extend the classification of standard planes to US video clips. This mimics what is currently done in clinical practice in which the clinician moves the probe until the plane is identified resulting in a 2D data stream.

A DL framework is proposed in [51] to detect FASP, FFASP and 4CH. To process the temporal information encoded in the US videos, the framework makes use of a long short-term memory (LSTM). Similarly, in [52], a classification CNN and a recurrent neural network (RNN) are used to detect FASP, TT, TC, and FLVSP.

2.3 Fetal standard-plane detection

To improve standard-plane detection performance and to increase the interpretability of the DL results, attention mechanisms, which endow CNNs to better focus on the most discriminating regions in the US image, have been exploited. The majority of the approaches in this regard implements the strategy shown in Fig. 2.4b. The work in [53] is among the firsts to add a self-gated soft-attention mechanism to a CNN for the detection of 13 US standard planes. Another popular solution in the field is to include sonographer gaze attention mechanism. In [54], a multi-task framework relying on SonoNet is proposed for FASP detection. The framework is trained to predict both standard plane and sonographer visual saliency prediction. In [55], a similar approach is proposed, which further processes US temporal clips via a temporal attention module. The considered planes are the FASP, FBSP and FFESP. Multiple data augmentation methods are exploited in [56] to improve standard plane classification.

Using classification CNNs may lead to inaccurate detection of standard planes, since it does not involve the detection of any specific anatomical landmark. This, in fact, does not reflect the way clinicians detect standard planes. A different approach investigated in the literature is to accomplish standard-plane detection through anatomical structure detection, as shown in Fig. 2.4c. This approach is followed in [57], where a classification CNN identifies images of the 4CH and a multi-task single-shot detector (SSD) detects the key anatomical structures of the plane as well as the US gain parameter and zoom of the image. In [58], a Faster R-CNN is used to assess the presence of fetal head specific anatomical structures in US images. A similar approach is followed in [59]. A multi-task framework, consisting of a Faster R-CNN with an additional classification branch, is used to detect 6 key anatomical structures and to evaluate if the head is centered in the image. A multi-task framework is also used in [60], in which a CNN inspired by Faster R-CNN is used to both predict the presence of specific anatomical structures in abdomen, brain and heart images and classify these structures. A hybrid approach is proposed in [61], where a CNN is trained to localise 13 anatomical structures by means of a weak supervision provided by image-level labels, thus, without the need for bounding-box annotation during training. In [62], TV detection is accomplished by localizing fetal brain by means of a segmentation CNN at first. Then, cavum septum pellucidum (CSP) visibility, fetal brain symmetry, and midline orientation are assessed by means of a number of additional CNNs.

More recently, semi-supervised and self-supervised strategies for standard-plane detection have been investigated. In [63], a semi-supervised pipeline is proposed to detect TT and TC from freehand fetal US video. The framework consists of a CNN for feature extraction, a prototypical learning module and a semantic transfer module to automatically label unseen video frames. In [64], a self-supervised method for scan plane detection in fetal 2D US images is proposed. Specifically, given an image, two small patches are randomly selected and swapped on the image and this procedure is repeated multiple times. A CNN is trained to restore the altered image

back to its original version. The CNN weights are then used to perform classification of several fetal standard planes (FASP, FBSP, FFESP, Kidney, Spine, 4CH, 3VV, RVOT, LVOT and facial profile). A semi-supervised learning approach to classify 13 standard planes is exploited in [65]. Authors' aim is also testing how background class, class imbalance and inter-class similarity influence performances.

A number of researchers is working to perform standard plane detection using deep reinforcement learning (DRL) from 3D fetal US, mainly to mimic what clinicians do and to explore inter-plane dependency. A common DRL scheme is shown in Fig. 2.4d. The work in [66] proposes a DRL to localize fetal brain standard planes in US volumes. The DRL framework is equipped with a landmark-aware alignment module that exploits a CNN to detect anatomical landmarks in the US volume. The landmarks are then registered to a plane-specific atlas. The DRL agent's interaction procedure is terminated by means of a RNN module. The approach is further enhanced in [67], which designs an adaptive RNN-based termination module to early stop the agent searching. A similar approach is performed in [68], which localizes multiple uterine standard planes in 3D simultaneously by multi-agent DRL. The latter is equipped with one-shot neural architecture search (NAS) module. To improve system robustness against the noisy environment, a landmark-aware alignment model is utilized. The spatial relationship among standard planes is learnt by a RNN.

US volumes are also processed in [69], where 2D image planes are fed to a CNN that predicts the 3D transformation to register each plane to a ground truth standard plane. In [70], a GAN is used to estimate the middle sagittal plane from 3D images; however, overfitting may arise considering that (low-dimensional) plane parameters are estimated from high-dimension volumetric data.

2.3.1 Limitations and open issues for fetal standard-plane detection

Methods used to perform fetal standard plane detection are heterogeneous. Even if classification algorithms have the advantages of being fast and light, they provide only rough localization information of the objects. Attention modules are thus added to better focus on important localization information rather than learning non-useful background information. Detection algorithms, on the contrary, better reflect what is actually performed during a scan plane detection, but with the drawbacks of being heavy and relying on bounding box, which may compromise doctor's understanding of the image.

From the review of the literature on standard-plane detection, a first limitation emerges in the number of considered planes. There are papers [51, 43, 44, 52] that consider only few planes. In some cases [54, 55, 42], only one plane is considered. When focusing on the experimental setup, only few papers refer to US scans on a patient basis. This information is, however, of crucial importance since US scans belonging to a woman should not be part of both training/validation and test sets to avoid bias. A second issue arises in terms of quantitative comparison among the

2.3 Fetal standard-plane detection

Table 2.6: Summary of deep-learning algorithms for anatomical structure analysis (for performance metrics refer to Table 2.2). *: Only the total number of subjects is reported. □: The dataset includes unhealthy subjects.

Paper (Year)	Organ	Training set size	Test set size	Annotators	Performance metrics	Gestational age
[73]	Heart	-	1991 2D	Few	$mAP=0.93$	-
[74]	Heart	- (10 sub- jects)	- (2 sub- jects)	Few	$Acc=0.82$	20 to 35 weeks
[75]	Heart	90 videos (12 sub- jects*)	1 video (12 subjects*)	Few	-	-
[76]	Heart	89 videos (10 sub- jects)	2 videos (2 subjects)	Few	$Acc=0.79$	20 to 35 weeks
[77]	Heart	350 videos (412 sub- jects*)	62 videos (412 sub- jects*)	1	$ACC=0.90$, $Prec=0.85$, $Rec=0.89$	≥ 28 weeks
[78]	Heart	413 2D (3 subjects*)□	106 2D (3 subjects*)□	Few	$IoU=0.94$, $Acc=0.96$	18 to 23 weeks
[79]	Heart	716 2D (895 sub- jects*)	179 2D (895 sub- jects*)	Few	$DSC=0.83$, $Acc=0.93$, $AUC=0.99$	-
[80]	Heart	1284 2D□	428 2D□	Few	$DSC=0.85$, $HD=3.33$, $Acc=0.93$	-
[81]	Heart	41 videos (40 frames each)	10 videos (40 frames each)	1	$HD=1.26$, $AD=0.20$, $DSC=0.94$	20 to 28 weeks
[82]	Heart	512 2D (319 sub- jects*)□	126 2D (319 sub- ject*)□	10	$DSC = 0.78$, $AP = 0.45$	-
[83]	Heart	624 2D□	69 2D□	1	Mean $DSC=0.83$	18 to 24 weeks
[84]	Heart	1033 2D (50 sub- jects*)	116 2D (50 subjects*)	1	$mAP=0.96$, $IoU=0.79$, $DSC=0.90$	18 to 24 weeks
[85]	Heart	- (100 subjects*)□	- (100 subjects*)□	Few	$Prec=0.85$, $Rec=0.86$, $F1=0.86$, $AUC=0.93$	-
[86]	Heart	410 2D (211 sub- jects*)	205 2D (211 sub- jects*)	Few	$IoU=0.55$, $DSC=0.68$	18 to 28 weeks
[87]	Heart	213 videos (363 sub- jects*)□	34 videos (363 sub- jects*)□	Few	$mAP=0.70$, $AUC=0.83$	18 to 34 weeks
[88]	Heart	107823 2D (1326 sub- jects*)□	4867591 2D (4666 sub- jects*)□	Few	$AUC=0.94$, $Rec=0.95$, $Spec=0.96$	18 to 24 weeks
[89]	Heart	2000 2D (2000 sub- jects*)□	100 2D (100 sub- jects*)□	2	$Prec=0.93$, $Rec=0.93$	-
[90]	Heart	3596 2D□	400 2D□	Few	$Acc = 0.85$	18 to 39 weeks
[91]	Heart	586 videos (minimum 80 and maximum 373 frames) (350 sub- jects*)	151 videos (minimum 80 and maximum 373 frames) (350 sub- jects*)	Few	$Acc=0.95$, $Rec=0.93$, $Spec=0.94$, $F1=0.95$	18 to 36 weeks

Chapter 2 Literature review on fetal-ultrasound analysis

Paper (Year)	Organ	Training set size	Test set size	Annotators	Performance metrics	Gestational age
[92]	Heart	89 videos (39556 2D) (12 subjects)	2 videos (12 subjects)	Few	-	20 to 35 weeks
[93]	Brain	-	4005 2D (1783 subjects)	Few	$DSC=0.77$, $IoU=0.63$, $HD=26.40$ pixel	28 to 40 weeks
[94]	Brain	-	448 2D (224 subjects)	3	$Prec=0.79$, $Rec=0.74$, $DSC=0.77$, $HD=0.78$	18 to 37 weeks
[95]	Brain	588 2D	146 2D	Few	$DSC=0.87$, $HD=28.15$, $Rec=0.86$, $Prec=0.90$	18 to 20 weeks
[96]	Brain	718 2D (HC18)	70 2D (HC18)	2	$DSC=0.97$, $Prec=0.97$, $Rec=0.98$, $HD=10.92$ mm	12 to 35 weeks
[97]	Brain	200 3D	45 3D	Few	$IoU=0.63$	20 to 29 weeks
[98]	Brain	271 3D	36 3D	Few	$DSC=0.82$, $BCE=0.09$	18 to 22 weeks
[99]	Brain	480 3D	48 3D	Few	$DSC=0.83$, $ED=2.24$ mm, $HD=4.51$ mm	20 to 25 weeks
[100]	Brain	138 3D	15 3D	Few	Surface Dice=0.92, Average Surface Distance=0.23 mm	20 to 26 weeks
[101]	Brain	50 3D (100 subjects*)	50 3D (100 subjects*)	2	$DSC=0.96$, $IoU=0.92$, $HD=0.46$ mm	20 to 31 weeks
[102]	Brain	Not specified (from 3D) [□] 599 2D	Not specified (from 3D) [□] 140 2D	Few	$Acc=0.99$, $ED=6.9$ mm, $IoU=0.82$, $HD=9.3$ mm	18 to 34 weeks
[103]	Brain	885 3D	300 3D	Few	$ED=1.36$ mm, $HD=9.05$ mm, $DSC=0.94$	14 to 31 weeks
[104]	Brain	8369 2D	869 2D	Few	$RMSE=13.85$, $MAE=16.05$, ≥ 14 weeks	
[105]	Brain	326 3D	121 3D	Few	Prediction error=6.9	18 to 36 weeks

2.3 Fetal standard-plane detection

Paper (Year)	Organ	Training set size	Test set size	Annotators	Performance metrics	Gestational age
[106]	Brain	689 3D	122 3D	Few	$MAE=4.1$, for Sylvian fissure; $MAE=5.1$, for Parieto-occipital fissure; $MAE=4.9$, for Calcarine fissure	19 to 30 weeks
[107]	Brain	29419 2D (12780 subjects*) [□]	4739 2D (12780 subjects*) [□]	15	$mAP=0.98$, $Rec=0.94$, $DSC=0.94$, $Acc=0.96$, $Spec=0.96$, $AUC=0.99$	22+4 weeks (normal) / 26+3 weeks (abnormal cases)
[108]	Placenta	1363 2D (247 sub-jects*)	205 2D (247 subjects*)	3	$DSC=0.92$, $Acc=0.93$	8 to 34 weeks
[109]	Placenta	10707 2D (321 sub-jects*)	2677 2D images (321 subjects*)	-	$Acc=0.81$, $Rec=0.88$, $Spec=0.65$	11 to 15 weeks
[110]	Placenta	27081 2D (67 sub-jects*)	5149 2D (67 subjects*)	1	$DSC=0.82$, $HD=31.84$ mm	19 to 33 weeks
[111]	Placenta	- (47 sub-jects*)	- (47 sub-jects*)	2	$DSC=0.86$	11 to 15 weeks
[112]	Placenta	- (70 subjects*) [□]	- (70 subjects*) [□]	-	$DSC=0.82$	17 to 37 weeks
[113]	Placenta	1196 3D (3104 sub-jects*)	1197 3D (3104 sub-jects*)	3	$DSC=0.84$, $HD=14.6$ mm	11 to 13 weeks
[114]	Placenta	280 3D (3064 sub-jects*)	20 3D (3064 subjects*)	Few	$DSC=0.73$, $HD=27$ mm	11 to 15 weeks
[115]	Placenta	30 3D (30 subjects*)	12 3D (30 subjects*)	-	$DSC=0.80$	24 to 34 weeks
[116]	Placenta	60 3D (104 subjects*)	44 3D (104 subjects*)	10	$DSC=0.64$, $HD=24.54$ mm	10 to 14 weeks
[117]	Placenta	99 3D (442 subjects*) [□]	25 3D (442 subjects*) [□]	-	$DSC=0.88$, $AUC=0.89$	11 to 14 weeks
[118]	Placenta	- 3D (60 patients*) [□]	- 3D (60 patients*) [□]	-	$DSC=0.76$	17 to 37 weeks
[119]	Amniotic Fluid	400 2D	900 2D	Few	$Acc=0.78$, $IoU=0.54$	22 weeks
[120]	Amniotic Fluid	310 2D (100 subjects)	125 2D (155 subjects)	1	$DSC=0.87$, $Prec=0.89$, $Rec=0.87$, $Spec=0.99$	20 to 36 weeks
[121]	Amniotic Fluid	2200 2D (1190 sub-jects*)	180 2D (1190 sub-jects*)	3	$DSC=0.86$, $Rec=0.81$, $Prec=0.93$, $HD=15.49$ mm	20 to 36 weeks
[122]	Lung	6312 2D	701 2D	-	$Acc=0.83$, $Rec=0.82$, $Spec=0.83$, $AUC=0.95$	20 to 41 weeks

Chapter 2 Literature review on fetal-ultrasound analysis

Paper (Year)	Organ	Training set size	Test set size	Annotators	Performance metrics	Gestational age
[123]	Lung	206 2D	126 2D	1	$MAE = 1.56$ mm	24 to 41 weeks
[124]	Kidney	40 3D	60 3D	2	$DSC=0.81$, $IoU=0.69$, $HD=8.96$ mm, 20 to 40 weeks	
[125]	Spine	320 3D	80 3D	-	-	15 to 41 weeks
[126]	Spine	-3D (3300 subjects*) [□]	-3D (3300 subjects*) [□]	2	$Rec=0.93$, $Prec=0.96$, $Acc=0.94$, $IoU=0.91$	23 to 31 weeks
[127]	Abdomen	126 3D	42 3D	-	$MAE=2.24$ mm	15 to 38 weeks
[128]	Abdomen	30 videos	3 videos	Few	AUC by Judd=0.87, $D_{KL} = 2.16$	-
[129]	Femur	30 3D (30 subjects)	20 3D (20 subjects)	1	$DSC=0.91$, $IoU=0.83$, $HD=4.08$ mm, $ED=0.87$ mm	23 to 31 weeks
[130]	Skull	52 3D	16 3D	Few	$DSC=0.83$, $IoU=0.70$	20 to 36 weeks
[131]	Head	10 3D (18 subjects*)	8 3D (18 subjects*)	1	$DSC=0.81$, $AUC=0.82$	Second trimester
[132]	Head	799 2D (HC18)	200 2D (HC18)	Few	$DSC=0.94$	12 to 35 weeks
[133]	Face	54 3D	6 3D	-	$ED=1.72$ mm	14 to 30 weeks
[134]	Face	120 3D	32 3D	-	$IoU=0.64$	-
[135]	Face	1658 2D	186 2D	3	$Acc=0.98$, $Rec=0.93$, $Spec=0.97$	19 to 38 weeks
[136]	Adipose Tissue	1370 2D	340 2D	3	$Rec=0.87$, $Spec=0.99$, $DSC=0.80$	17 to 41 weeks
[137]	Adipose Tissue	65 2D	50 2D	-	$Prec=0.92$, $Rec=0.94$, $Spec=0.95$, $DSC=0.93$	26 to 41 weeks
[138]	Multiple	3109 2D	965 2D	1	$Acc=0.97$, $Prec=0.76$, $Rec=0.75$	18 to 20 weeks
[139]	Body, Head, Leg	10000 2D	2000 2D	1	$Rec=0.92\%$	16 to 34 weeks
[140]	Multiple	31629 2D	3986 2D	Few	Top-1 $ACC=0.77$, Top-3 $ACC=0.94$	-

2.3 Fetal standard-plane detection

Paper (Year)	Organ	Training set size	Test set size	Annotators	Performance metrics	Gestational age
[141]	Abdomen, Head, Heart, Spine	12808 2D	9979 2D	-	$Prec=0.96$, $Rec=0.96$, $DSC=0.96$	-
[142]	Abdomen, Heart, Skull	225825 (free hand) and 30048 (single sweep) 2D	7512 (single sweep) 2D	-	$Acc=0.73$ and $Acc=0.89$ for 3- and 4-class classification, respectively	-
[143]	Kidney, Liver, Spleen	149775 2D	37444 2D	2	$Acc=0.85$	-
[144]	Abdomen, Heart, Skull	365 videos	91 videos	-	$AP=0.91$	-
[145]	Abdomen, Head	1588 2D	741 2D	1	$DSC=0.97$, $IoU=0.96$	19 to 40 weeks
[146]	Abdomen, Head	4200 2D (Device A)	1516 2D (Device B), 816 2D (Device C)	Few	$DSC=0.93$, $HD=12.50$, $IoU=0.89$, $Rec=0.92$, $Prec=0.95$ for abdomen, $DSC=0.97$, $HD=6.07$, $IoU=0.94$, $Rec=0.97$, $Prec=0.96$ for head	-
[147]	Fetus, Gestational sac, Placenta	60 3D	44 3D	10	$DSC=0.80$, $HD=14.13$ mm	10 to 14 weeks

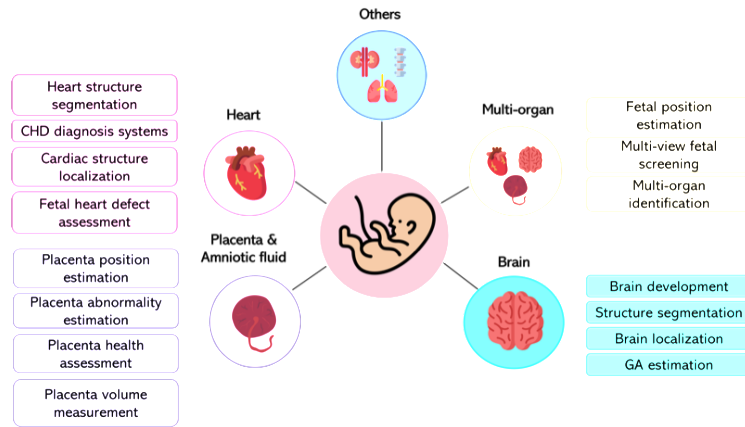


Figure 2.5: Overview of the most common tasks in the literature for fetal organs analysis.

different approaches. The use of different datasets, some of which are really small in size, does not allow making a fair comparison among the solutions proposed in the literature. An important step towards a fair evaluation of methods has been done in [39], which released the first dataset in the field. However, only FASP, FBSP, 4CH, maternal cervix, FFESP are considered. In addition, the majority of the surveyed papers neglects the temporal dimension and treats each image individually. This may be considered as an issue for standard-plane detection, as clinicians identify the standard plane by manoeuvring the US probe to the desired scan plane, which results in a 2D data stream. Showing training/validation curves as well as using visual explanation techniques, e.g. Grad-CAM [148], should also be considered for a fair assessment of algorithm performance, even in terms of model bias and variance.

2.4 Anatomical structure analysis

Anatomical structure analysis is one of the most important part of fetus screening, as it provides a direct evidence of possible fetal malformations, as well as placenta localization disorders and premature birth risk. According to ISUOG guidelines [149], from 18-22 weeks of pregnancy organs are already developed. However, organ evaluation during the first trimester (from 11 to 13+6 weeks) could also be performed to early detect fetal major anomalies. Fetal screening during mid-trimester involves brain, face, neck, chest/heart, abdomen, skeletal, placenta, umbilical cord, genitalia and cervix.

This section surveys methods for the analysis of anatomical structures, including fetal heart (Sec. 2.4.1), brain (Sec. 2.4.2), and placenta and amniotic fluid (Sec. 2.4.3), for which a relatively rich literature already exists. DL approaches for the analysis of other fetal anatomical structures are grouped in Sec. 2.4.4, while Sec. 2.4.5 surveys approaches for multi-structure analysis. Figure 2.5 summarizes the main tasks addressed in the literature of anatomical structure analysis and Table 2.6 lists

all the papers surveyed in this section.

2.4.1 Heart

Fetal cardiac evaluation is of crucial importance to detect heart diseases, such as congenital heart diseases (CHDs) and intrauterine growth restriction. Cardiac evaluation generally consists of cardiac-function analysis and heart anatomical evaluation, including heart dimension and shape.

A number of DL approaches in the field focuses on heart and heart-structure detection.

The work in [73] uses a SSD model with aggregated residual visual blocks to detect anatomical-heart structures such as left atrial pulmonary vein angle, apex cordis, moderator band, and multiple ribs in 4CH. A cardiac-structure localization algorithm is proposed in [74]. The presence of the heart in the 4CH is detected using a modified VGG-16, then a Faster RCNN model coupled with LSTM layers is used to temporally classify the presence of Foramen ovale, Mitral valve, Tricuspid valve, LV wall and RV wall. In [75], presence, viewing plane, location and orientation of the fetal heart is predicted by formulating the problem as a multi-task prediction within a recurrent CNN detection framework.

In [76], the same task as in [75] is formulated as a multi-task learning problem within a hierarchical convolutional model that progressively encodes temporal information throughout the network. In [77], spatio-temporal representations of fetal heart are learned by means of an end-to-end two-stream fully CNN for temporal sequence analysis. The goal is to capture motion and appearance features in a weakly supervised manner.

Structure segmentation offers more information than detection, since heart and heart-structure shape is a significant indicator of possible pathologies. Here, encoder-decoder CNNs are often exploited in the literature. In [78], a U-Net architecture is used to segment fetal cardiac standard planes to early detect possible heart structural abnormalities. A cascaded network is developed in [79] to accurately segment 7 anatomical structures in the 4CH view. The network consists of a dilated sub-network responsible for aggregating both global and local information and two stacked U-Nets. A cascaded U-Net is also used in [80] to segment anatomical heart structures. Segmentation of fetal left ventricle from fetal US videos is proposed in [81]. A dynamic CNN trained with multiscale information and different fine tuning strategies is used. The first cardiac frame is labelled to fine-tune the CNN. As the frames are segmented sequentially, the CNN is fine-tuned dynamically by shallow tuning to fix the latest frame. To differentiate the left ventricle and atrium regions, the mitral valve base points are tracked.

Instance segmentation of heart structures is performed in a number of papers, including [82] where the four cardiac chambers are segmented with a network with three branches: the category branch, the mask branch, and a category-attention

branch. This latter is used to correct the instance mis-classification and improve the segmentation accuracy. A Mask-RCNN is used in [83] to detect and segment the left and right atria, left and right ventricles, aorta and hole. The proposed approach is used to assess fetal heart defects (atrial septal, ventricular septal and atrioventricular septal defects). A Mask RCNN is used in [84] to segment the fetal heart in 4 different standard views (3VT, 4CH, LVOT, RVOT) and fetal heart chambers in each view to search for possible heart defects.

CHD diagnosis is another popular task investigated in the literature of fetal heart analysis. A hypoplastic left heart syndrome (HLHS) detector is developed in [85]. The detector consists of a SonoNet to detect standard planes (4CH, LVOT, RVOT) and a VGG16 to identify HLHS patients versus healthy subjects. In [86], ventricular septal defects are evaluated by means of YOLOv2, which detects the ventricular septum, and U-Net, that is used to segment the cropped ventricular septum area. A calibration module is used to further enhance U-Net output. YOLOv2 is also used in [87] for the detection of abnormalities in cardiac substructures from 4CH and 3VT 2D frames extracted from videos. In [88], an ensemble of deep residual networks are used to classify five heart views and successively classify normal hearts and complex CHD. A modified U-Net is further trained to calculate cardiothoracic parameters such as the cardiothoracic ratio (CTR), cardiac axis (CA) and fractional area change (FAC) for each cardiac chamber. In [89], a CNN with a residual learning module linked with guided back-propagation to visualize feature maps is used to diagnose fetal CHD. In [90], a one-class classification network is proposed to classify patients with CHD and healthy subjects. A GAN is used for data augmentation and compared with state-of-the-art approaches. In [91], a YOLOv3 is exploited to detect the four-chamber views at first. Then a mobileNet is used to classify samples into end diastole or end systole. In order to tackle lack of available data in fetal CHD domain, the work in [92] proposes an incremental learning approach to build a hierarchical network model that allows for a parallel inclusion of previously unseen anatomical classes without requiring prior data distributions. The goal here is the detection of different anatomical structure in different fetal cardiac views. The pipeline relies on natural hierarchies in US videos and it is built to account for new data in a self-organized fashion.

2.4.2 Brain

Assessing fetal brain development is crucial to evaluate fetus growth and diagnose brain pathologies. In clinical practice, several structures are analyzed to assess the fetal brain development and well-being. From the clinician side, this is not effortless due to challenges such as the high intra- and inter-structure variability. Automatic fetal brain analysis commonly includes anatomical structure localization, segmentation, classification, and measurement. Structures considered in the literature include middle cerebral artery, CSP, cerebellum, brainstem, ventricles, thalamus,

cortical plates and fissures (Sylvian, Parieto-occipital, calcarine).

Among DL approaches for fetal brain analysis, a number of work focuses on brain-structure segmentation using encoder-decoder architectures. The work in [93] proposes a DL-based architecture to segment the middle cerebral artery and provide the gate position on Doppler US images to the sonographer. The architecture uses a pre-trained dilated residual network as encoder and dense upsampling convolution blocks as decoder. After the segmentation, the gate position is retrieved as the center of the segmented area. The work in [94] proposes a deep attention network inspired by the U-Net encoder-decoder architecture to segment the CSP and measure its width to evaluate the presence of anomalies. VGG11 is used as backbone in the encoding path and a channel attention module is introduced between the encoder and the decoder. Similarly, a U-Net inspired CNN (i.e., ResU-Net) is used in [95], with the aim to automatically segment the cerebellum. ResU-Net encodes residual blocks and dilated convolutions to recover the spatial resolution lost in the encoder keeping the number of training parameters low. In [96], a U-Net inspired CNN (i.e., MA-Net) is proposed for fetal head circumference segmentation. MA-Net is based on an encoder-decoder architecture consisting of 5 modules: encoder, atrous convolution, pyramid pooling, decoder and residual skip pathway modules.

More recently, a number of segmentation approaches has been exploring 3D architectures for processing the 3D information naturally encoded in 3D US data. A 3D U-Net is used in [97] to mask out extra-cranial tissues as a preliminary step to simultaneously segment and localize 5 brain structures (lateral ventricles, CSP, thalamus, cerebellum, and cisterna magna). The output of the 3D U-Net is projected in three standard planes (axial, sagittal and coronal), which are then processed by three CNNs, one for each standard plane. Each CNN outputs the 2D segmentation mask of the 5 structures. A 3D U-Net is also used in [98], [99], [100] and [101]. In [98], it is used to segment the fetal cortical plate and measure the depth of the Sylvian fissure on a dataset annotated by expert clinicians using an atlas. Similarly, in [99] a multi-task CNN based on 3D U-Net is used for the automatic segmentation of the white matter, thalamus, brainstem, and cerebellum. In [100], the authors propose a 3D U-Net combined with active contours for CSP segmentation. In [101], a 3D U-Net is trained to segment the whole fetal head. The 3D U-Net is combined with a hybrid attention scheme to enhance the feature maps.

Brain-structure segmentation is further explored and included in a multi task context in [102]. The authors propose an approach to align 3D US brain volumes, recovered from a stack of axial slices, to a coordinate system based on skull boundaries, eye socket location, and head pose. A multi-task FCN is trained to address the problem of 3D fetal brain localization, structural segmentation, and alignment. A method for brain localization is also proposed in [103], which explores the use of an end-to-end 3D CNN for automated brain localization and extraction from 3D fetal US. Differently from [102], which predicts the position of the brain from 2D slices extracted from 3D volumes, this is a fully 3D approach and relies on a modified 3D

U-Net for brain extraction.

Other than structure segmentation and localization, researchers are also working on GA estimation and brain development analysis. An approach to GA estimation from 2D trans-thalamic US images is proposed in [104]. The authors propose a Bayesian Neural Network (BNN) with a VGG-16 backbone and an auxiliary regression model, which are trained to predict calibrated aleatoric and epistemic uncertainties on GA. Methods to evaluate brain development are investigated in [105] and [106]. The work in [105] presents a model to predict brain maturation from 3D US through the use of a 3D Convolutional Regression Network. In [106], the authors propose a method to estimate the development of three brain fissures (i.e., Sylvian, Parieto-occipital and Calcarine) from 3D US volumes, by predicting the fetal GA based on their respective morphology. The regions relative to each fissure are extracted from the US volumes and passed to three separate ResNet to predict the GA of each region.

As for CAD of brain pathology, in [107] the authors propose a DL-based pipeline to support the diagnosis of brain lesion from 2D images. The pipeline relies on three architectures for (i) cranio-cerebral segmentation, using U-Net with dilated convolutional, (ii) image classification in normal/ abnormal, using VGG-19 pre-trained on ImageNet, and (iii) lesion localization, using Grad-CAM. The lesion localization, in the form of Grad-CAM activation heatmaps, is visually evaluated by a doctor.

2.4.3 Placenta & amniotic fluid

The placenta is an organ that provides oxygen and nutrients to the fetus, ensures thermo-regulation, and removes waste products from the fetus's blood. This structure is closely related to fetus health: abnormal placental function may affect the development of the fetus, and in severe cases, even endanger the life of the fetus. Given its relevance, several clinical assessments are performed to evaluate the placenta condition along with the fetus health. In this scenario, placenta segmentation may provide automatic quantification of placenta volume and morphology.

A number of approaches performs placenta segmentation from 2D US images. The approach proposed in [108] uses a U-Net-inspired CNN. The CNN is modified adding a layer to detect acoustic shadow and improve the segmentation accuracy. Similarly, in [109] a U-Net-based CNN is used for placenta segmentation. An EfficientNet is then used for the classification of normal or abnormal placenta. A U-Net-like architecture for placenta segmentation is also used in [110]. The authors introduce an auxiliary classification task, incorporating the prediction of the placental position (anterior or posterior) in the U-Net architecture and improve the segmentation accuracy. In [111], three 2D cross-sectional images manually extracted from 3D US, are used to train an encoder-decoder CNN architecture. An atlas-based joint label fusion algorithm is then applied to the CNN output to combine the three predictions and enhance segmentation performance. Predictions from CNN and joint label fusion

are combined via a random forest model to obtain the final segmentation.

In [112], stacked GANs are used to automatically segment placenta along with its peripheral vasculature, as well as acoustic shadows, with the goal to provide surgeons with context awareness during twin-to-twin transfusion syndrome (TTTS) treatment.

Other researchers are working on placenta segmentation from 3D US data. The work in [113] uses a 3D U-Net.

For the same goal, in [114] the authors use an open-source CNN trained with labels obtained through a semi-automatic Random Walker strategy.

The work in [115] proposes a three stage pipeline for whole placenta segmentation and volume estimation from multi-view 3D US volumes. The pipeline includes: (i) multi-view acquisition through a multiplexer device that fixes 3 probes in an angle of 30 degrees to each other and switches the view from one probe to the other almost in real time; (ii) voxel-wise fusion to combine the multiple views; (iii) segmentation based on a 3D U-Net. The work in [116] proposes an approach to automatically segment placenta, fetus and gestational sac. The first step includes a customized 3D CNN with long skip connections to segment the 3 structures. A RNN is used to add contextual knowledge and refine the semantic segmentation. Finally, a hierarchical deep supervision mechanism is used to boost the flow of information in the RNN and improve the segmentation performance.

Volumes of 3D US obtained from singleton pregnancies were automatically segmented by [117]: their CNN pipeline combines two 2D and one 3D models with downsampling/upsampling architectures. The placental volumes, derived from the automatic segmentation model, are used to train multivariable logistic-regression classifiers to predict small-for-gestational-age infants born under the 10th/5th centiles (SGA10/SGA5).

An interesting approach to placenta segmentation from multimodal imaging (i.e., 3D US and magnetic resonance imaging) is proposed in [118]. The approach combines radiomics and DL to efficiently segment intra-uterine tissues.

Amniotic fluid plays an important role in fetal well-being and development. Amniotic fluid has a myriad of functions: it protects fetus and umbilical cord, prevents infections and provides the necessary growth factors to allow normal development and growth of fetal organs. In the clinical practice, to assess the sufficiency of amniotic fluid quantity, the amniotic fluid index is used. This index is calculated dividing the maternal abdomen into 4 quadrants using the midline and the umbilicus, then the deepest pocket of amniotic fluid is evaluated in each quadrant. The index is given by the sum of the 4 measurements. With a view to automatize the index computation, amniotic fluid segmentation is a crucial task.

In [119], an encoder-decoder architecture with VGG16 as backbone is used to segment amniotic fluid and fetal body in 2D US images. In [120], a 2-step framework is used to segment the amniotic fluid pocket and measure the amniotic fluid index. The segmentation step is performed with a modified U-Net architecture, called AF-

net, which combines atrous convolution and multi-scale side-input and side-output layers. This approach is further improved in [121], which proposes a dual-path network to segment the amniotic fluid volume from 2D US images. The primary path consists of AF-net, while the secondary path is an auxiliary CNN used to remove reverberation artifacts and complement the primary path prediction. The final segmentation output is obtained combining results of primary and secondary paths.

2.4.4 Others

Lungs

Immaturity of fetal lung development is the primarily cause of neonatal respiratory morbidity.

Quantitative US imaging is often used as a non-invasive tool for fetal lung maturity assessment, through US visualization at 4CH level. The normal fetal lung has very similar echogenicity to the adjacent liver, but with slightly different texture; and in presence of abnormalities either increasing or decreasing in echogenicity can be seen.

Recently, DL approaches have been introduced to tackle the challenges of lung analysis from fetal US image, especially for estimating GA.

The framework in [122] is based on DenseNet, which is developed for processing US images of 4CH plane and to predict the GAs based on 3 classes: class I (from 20 to 29+6 weeks), class II (from 30 to 36+6 weeks) and class III (from 37 to 41+6 weeks).

For the same aim, the work in [123] proposes a two-stage transfer learning approach. A U-Net-like architecture with residual connection is first trained to recognize samples of fetal lung regions from other regions taken from the US images; then, the pre-trained U-Net is tuned on fetal lung region samples only to regress the corresponding GA, and thus its maturation degree. The study is performed on 4CH US scans, manually segmenting the lung region.

However, this work provides only an indirect prediction of the lung maturation degree, as it relies on the direct correlation between the lung maturity and the GA, which does not always exist.

Kidney

In-utero assessment of kidney is crucial to perform early diagnosis of renal pathologies. Poor kidney development is known to be associated with increased risk of kidney disease into adulthood. At the same time, abnormal fetal growth is associated to a reduced functionality of kidneys after birth [150, 151].

In this field, the work in [124] proposes a 3D U-Net to perform kidney segmentation from 3D B-mode and Power Doppler volumes. Despite the small number of testing volumes, this is an interesting study that may pave the way for the development of

early diagnosis tools for fetal kidney.

Spine

Fetal spine length provides insights into the fetal growth as it is affected by a variety of malformations (spina bifida, meningocele, diastematomyelia, vertebral segmentation anomalies, sacral agenesis, spinal dysgenesis, spondylothoracic or spondylocostal dysplasia). To provide an assessment of fetal spine, the work in [125] focuses on the identification of spine centerline from 3D US scans. Spine segmentation is performed with a CNN that processes images at multiple scales and different fields of view. The prediction is used as input for a model-based tracing algorithm responsible to draw spine centerline.

For a similar aim, the work in [126] presents an approach to identify the spina-bifida and segment it. The approach is based on an efficient implementation of U-Net replacing standard convolution with the octave convolution proposed in [152].

This approach achieves high recognition accuracy, good segmentation accuracy, and short running time, but it is limited on few cases, without dividing among different spina bifida types and with an imbalanced presence on malformation cases compared to healthy ones.

Abdomen

Fetal abdomen assessment is performed to evaluate important prognostic parameters of neonatal morbidity and mortality and to assess fetal growth. Segmentation of the abdominal outline is of interest for measuring fetal abdominal circumference. For this purpose, the work in [127] develops an approach based on CNNs to extract image features to be integrated in a deformable model. Fetal abdomen localization is achieved in [128] using an encoder-decoder architecture for US video saliency prediction. The encoder part consists of a truncated SonoNet, while the spatio-temporal decoder of the network is made up of a bidirectional gated-recurrent-unit recurrent convolutional network (GRU-RCN).

Femur

Volume and length of fetal femur have unique importance in fetal weight estimation. However, it is hard to be evaluated through US, due to the difficulty in locating the tips of the femur, boundary deficiency and ambiguity for tissues' low contrasts, and variation of pose, shape and size of this structure. In [129], the authors develop a unified framework for simultaneous segmentation and landmark localization of fetal femur in prenatal US volumes: fetal femur ROI is first identified through a U-Net model, then segmentation and landmark localization branches receive the common features of ROI extracted by the shared layers and generate task-specific descriptors.

Chapter 2 Literature review on fetal-ultrasound analysis

Structure segmentation	CAD systems	Gestational age	Structure detection	Fetal weight	Structure classification
 • 2D  • 2D  • 3D	 • 2D  • temp	 • 2D  • 3D	 • 2D  • temp	 • 3D	 • 2D  • temp
 • 2D  • 3D	 • 2D  • 3D	 • 2D  • 3D	 • 2D  • 3D		
 • 2D  • 3D	 • 3D				

Figure 2.6: Most addressed tasks (from left to right) in the field of anatomical structure analysis. Lilac = heart, cyan = brain, purple = placenta and amniotic fluid, yellow = multi-organ, blue = others, temp = temporal information.

Skull and Head

The evaluation of fetal head is a critical part of sonographic examination. This evaluation is dependent on operator experience and US image intrinsic characteristics. To attenuate these issues, DL algorithms in the literature mainly focus on the segmentation of the fetal head. The work in [130] proposes a two-stage approach for skull segmentation in fetal 3D US. A 3D U-Net is used to roughly segment the skull. This segmentation concatenated with two additional channels (i.e., the US-wave incidence angle map and the US shadow casting map) and fed to a second 3D U-Net. The work in [131] aims at merging several partially occluded US volumes, acquired by placing the US transducer at different projections of the fetal head, to compound a new US volume containing the whole brain anatomy. For this aim, the authors propose a pipeline of 4 CNNs. The first 2 CNNs follow what is done in [131], while the additional 2 CNNs perform US-volume registration. A different approach is proposed in [132], with the aim of segmenting fetal head. A CycleGAN is trained with a set of unpaired images and auxiliary masks obtained from a shape prior model, to generate the pseudo-labels corresponding to each of the training images. The CycleGAN is equipped with a Variational Auto-encoder based discriminator and a Discriminator-guided Generator Channel Calibration module which calibrates the pseudo-label generator using the discriminator's feedback to improve the pseudo-labels.

Face

Accurate detection and visualization of fetal face position and orientation is crucial in prenatal diagnosis, growth monitoring and detection of fetal anomalies. A U-Net inspired encoder-decoder 3D CNN is proposed in [133] to segment fetal face from 3D US volumes. In [134], an RPN-based object detection framework is proposed to detect landmarks in 3D facial US volumes. Predictions from the RPN architecture are further refined with a distance-based graph prior to produce the final bounding box for each landmark. Fetal facial expression is another important aspect to take into consideration for the evaluation of the fetal brain and central nervous system.

In [135], a custom built CNN classifier is trained to recognize 7 fetal expressions in 2D US images.

Adipose tissue

Observation and evaluation on fetal adipose tissue is significant to determine the growth and nutritional adequacy of the fetus.

The work in [136] performs fetal adipose tissue segmentation through a U-Net with depth-wise separable convolutions. The work in [137] proposes a framework for detecting and segmenting the fetal thigh cross-sectional area of adipose tissue. The framework relies on Faster R-CNN for localizing the regions containing the fetal thigh cross-section, on which threshold-based segmentation is performed to allow measurement of adipose tissue thickness.

2.4.5 Multi-organ analysis

DL researchers are working on multi-organ analysis to reproduce the actual clinical screening, which accounts for a comprehensive list of anatomical structures.

A number of DL learning algorithms in the field focuses on classification tasks. The work in [138] proposes a method to automatically classify 14 different fetal structures (abdomen, arm, blood vessels, cord insertion, face, femur and humerus, foot, genitals, head, heart, kidney, leg, spine, hand) from US images by fusing information from both cropped regions of fetal structures and whole images. Two CNNs pre-trained on ImageNet are used as feature extractors. The features are classified by means of support vector machines. Similarly, the work in [139] fine-tunes a VGG16 to classify US images in four classes (head, body, leg and other). This is needed as preliminary step towards the estimation of fetal position. Based on the output of Grad-CAM, the body parts position in the images is estimated. Spatio-temporal classification is performed in [140], in which a LSTM is used to classify 11 anatomical categories (heart, background, brain with skull and neck, Doppler maternal anatomy, spine, abdomen, nose and lips, kidneys, face side profile, femur, other).

In [141], the authors propose a method for multi-organ (abdomen, head, heart, spine) classification from the integration of fetal US images with corresponding textual descriptions. The authors exploit a curriculum learning approach to train a NLP-based fetal US image captioning model with a dataset prepared using real-world US videos along with synchronized and transcribed sonographer speech recordings: 12808 and 9979 image-caption pairs are used for training and testing sessions, respectively.

An unsupervised approach to multi-organ classification is proposed in [142], which develops a cross-device and cross-anatomy adaptation network to classify heart, abdomen and skull of an unlabelled single-sweep video dataset guided by knowledge of a labelled free-hand scanning protocol video dataset.

The network consists of encoder, projection layer, anatomy classifier, domain classifier and two mutual information discriminators.

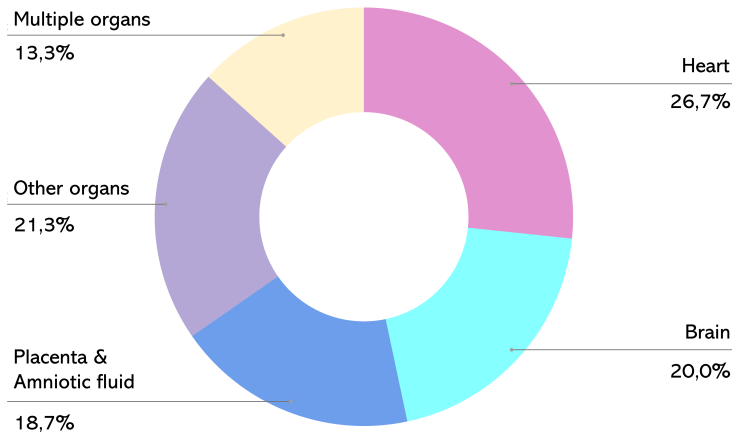


Figure 2.7: Number of papers surveyed in Sec. 2.4 according to the anatomical structure.

A different approach is used in [143], where a multi-task learning framework is proposed to classify 11 different views related to abdominal organs and detect 14 landmarks from kidney, liver and spleen. This framework relies on a shared ResNet encoder and two branches, for classification and landmark detection.

Multi-organ analysis also involves DL algorithms for localization and segmentation tasks. In a weakly supervised fashion, the work in [144] proposes a model consisting in a CNN for extracting features and an attention-gated LSTM to localize skull, abdomen and heart in consecutive frames, considering also non-standard planes. A pipeline to segment fetal head and abdomen is proposed in [145]. The pipeline consists of three cascaded FCN and each FCN take as input the summation of fetal US image with prediction map from the previous level to refine boundaries of the anatomy segmentation.

In [146], segmentation of fetal head and abdomen is performed based on style transfer, a technique used to extract the semantic content from a target image and render it to a source image. The authors use this technique to tackle the problem of appearance shift: a U-Net is combined with style transfer based on dynamic instance normalization to perform organ segmentation from US images acquired with three different devices.

In [147], a framework to simultaneously segment fetus, gestational sac, and placenta is proposed. The framework consists of a 3D CNN, which explores spatial intensity concurrency, and a RNN, which encodes spatial sequential to improve boundary refinement.

2.4 Anatomical structure analysis

Table 2.7: Summary of deep-learning algorithms for biometry parameter estimation (for performance metrics refer to Table 2.2). HC: Head circumference, BPD: Biparietal diameter, OFD: occipito-frontal diameter, AC: Abdominal circumference, FL: Femur length, TCD: trans-cerebellar diameter. *: Only the total number of subjects is reported. □: The dataset includes unhealthy subjects.

Paper	Biometry	Training set size	Test set size	Annotators	Performance metrics	Gestational age
[153]	HC, BPD	2164 2D	100 2D	2	$DSC=0.98$, $MAE=1.3$ mm	18 to 22 weeks
[154]	HC	999 (HC18)	2D 335 (HC18)	2D 1	$DSC=0.95$, $MAE=2.44$ mm	12 to 35 weeks
[155]	HC	999 (HC18)	2D 297 (HC18)	2D 1	$DSC = 0.97$, $MAE = 1.89$ mm	12 to 35 weeks
[156]	HC	999 (HC18)	2D 335 (HC18)	2D 1	$DSC=0.98$, $MAE=1.77$ mm	12 to 35 weeks
[157]	HC	999 (HC18)	2D 335 (HC18)	2D Few	$DSC=0.76$, average error = 14.96%	12 to 35 weeks
[158]	HC	849 (HC18)	2D 150 (HC18)	2D 1	$DSC=0.97$, $MAE=2.27$ mm	12 to 35 weeks
[159]	HC	HC18	HC18	1	$DSC = 0.95$, $HD=4.5$ mm	12 to 35 weeks
[160]	HC	HC18	HC18	1	$MAE=2.16$ mm	12 to 35 weeks
[161]	HC	HC18	HC18	1	$DSC = 0.97$, $MAE=2.12$ mm	12 to 35 weeks
[162]	HC	849 (HC18)	2D 150 (HC18)	2D 1	$DSC=0.92$, $MAE = 2.22$ mm	12 to 35 weeks
[163]	HC, BPD, OFD	999 (HC18)	2D 335 (HC18)	2D 1	$DSC = 0.97$, $MAE = 1.81$ mm	12 to 35 weeks
[164]	HC, BPD	102 2D	70 2D	2	$ACC=0.87$, success rate = 0.93	18 to 35 weeks
[165]	HC	999 (HC18)	2D 335 (HC18)	2D 1	$DSC = 0.97$, $MAE = 2.33$ mm	12 to 35 weeks
[166]	HC	999 (HC18)	2D 335 (HC18)	2D 1	$DSC = 0.97$, $MAE = 1.90$ mm	12 to 35 weeks
[167]	HC	999 (HC18)	2D 335 (HC18)	2D 1	$DSC = 0.98$, $MAE = 1.95$ mm	12 to 35 weeks
[168]	HC	905 (HC18)	2D 94 2D (HC18)	1	$DSC = 0.97$	12 to 35 weeks
[169]	HC	2848 (2000 subjects*)	2D sub-540 2D (2000 subjects*)	45	$DSC = 0.98$, $MAE = 1.81$ mm	12 to 22 weeks

Chapter 2 Literature review on fetal-ultrasound analysis

Paper (Year)	Biometry	Training set size	Test set size	Annotators	Performance metrics	Gestational age
[170]	HC	800 2D (HC18)	199 2D (HC18)	1	$MAE=4.52$ mm	12 to 35 weeks
[171]	HC	800 2D (HC18)	199 2D (HC18)	1	$MAE = 4.78$ mm	12 to 35 weeks
[172]	AC	56 2D (1 subject*)	32 2D (1 subject*)	2	$DSC=0.85$, $ACC=0.79$	20 to 34 weeks
[173]	AC	- (77 sub-jects*)	- (77 sub-jects*)	-	$DSC=0.92$, $ACC=0.87$	-
[174]	CRL	697 2D	545 2D	1	$DSC=0.93$, $IoU=0.88$, $MAE=3.87$ mm	11 to 15 weeks
[175]	LV	4379 2D [□]	500 2D [□]	3	$Acc=0.96$, $Prec=0.98$, $Spec=0.90$, $MAE=1.80$ mm	-
[176]	FL	2300 2D (435 sub-jects*)	310 2D (435 subjects*)	1	$DSC=0.92$, $MAE=0.46$ mm	-
[177]	CTR	1669 2D	417 2D	2	$DSC=0.93$	18 to 39 weeks
[178]	HC, AC, FL	262 2D (42 subjects*)	87 2D (42 subjects*)	1	$MAE=2.67$ mm (HC), $MAE=3.77$ mm (AC), $MAE=2.10$ mm (FL)	-
[179]	HC, AC, FL	274275 2D (560 patients)	57001 2D (140 patients)	6	$DSC = 0.96$, $MAE=2.9$ mm (HC), $MAE=3.8$ mm (AC), $MAE=0.8$ mm (FL)	15 to 38 weeks
[180]	CRL, HC, AC	44 3D	21 3D	Few	$IoU=0.92$ for abdomen and head, $IoU=0.66$ for limbs segmentation, $MAE=2.24$ mm	11 to 14 weeks
[181]	AC, BPD, FL, HC	999 2D (HC18) + 1154 2D (551 sub-jects*)	335 2D (HC18) + 180 2D (551 subjects*)	2	$DSC=0.98$, $HD=1.14$ mm, average perpendicular distance=0.2 mm	12 to 35 weeks
[182]	HC, TCD	937 2D (937 subjects)	913 2D (913 subjects)	1	$MAE=2.04$ mm	-
[183]	BPD, HC, CRL, AC, FL	147855 2D (4433 sub-jects)	7233 2D (2491 sub-jects)	Few	$Acc=0.93$, $IoU=0.91$, $MAE=1.89$ cm	13 to 18 weeks
[184]	HC, BPD	30000 2D	1000 videos	1	$Acc=0.96$	18 to 42 weeks

2.4.6 Limitations and open issues for anatomical structure analysis

As shown in Fig. 2.6, segmentation is the most addressed task for all the anatomical structures. However, most of the proposed methodologies consist of 2D U-Net-based architectures, while more advanced architectures exploited in closer fields (e.g. adversarial segmentation, spatio-temporal processing) are less investigated. Interestingly, priors relevant to anatomical structure shape have not been fully exploited yet. However, shape-constraint DL strategies have been recently proposed in the literature to drive the output of segmentation CNN towards the desired structure shape, for example by using adversarial learning [185, 186].

Up to now, the most studied anatomical structures are heart and brain, contributing with 26.7% and 20.0% of the surveyed papers, respectively, as shown in Fig. 2.7.

As regards fetal heart analysis, papers in literature deal with the most common screening examinations such as 4CH and LVOT/RVOT views assessment, ventricular and atrial chambers evaluation and fetal heart location and orientation estimation along with fetal heart motion tracking. This reflects what is currently considered by ISUOG guidelines to be the best fetal heart evaluation practice [149]. However, at present such assessments are performed as distinct tasks, trained and evaluated on different datasets. The heterogeneity of these approaches doesn't allow comparison among algorithms: this is particular evident in CHD assessment, which is evaluated considering different approaches (from cardiac structure segmentation to global image classification).

In addition, for a better functional and anatomical assessment of the fetal heart, volume sonography (3D/4D spatiotemporal image correlation, STIC) and tissue Doppler can also be used, as declared in the ISUOG guidelines [149]. However, these techniques have not been exploited yet in the domain of DL for fetal image analysis.

Applications for DL in fetal brain anatomical structure analysis are heterogeneous as well and each one focuses on a particular task, such as brain localization, segmentation, classification or GA estimation. Given the fragmented nature of the literature available in this field, comparing the different approaches is challenging. The lack of public datasets further hampers a fair comparison among the different approaches.

Despite the variety of structures that are evaluated by clinicians during fetal screening, only part of these structures has been considered in the literature so far (Fig. 2.7). More importantly, most of the work does not focus on detecting problematic pregnancies, which is the main goal of fetal screening, but is limited to anatomical structure detection or segmentation. Some groups are working in this direction (e.g., [117, 88, 107] for detecting small for GA infants, CHD, and abnormal brain, respectively), but there are still open research questions.

The same considerations can be applied to fetal placenta evaluation domain, whose morphology, size and location are estimated in order to determine possible malformations.

Only 2 papers out of 11 clearly include twin pregnancies (mono- and dichorionic

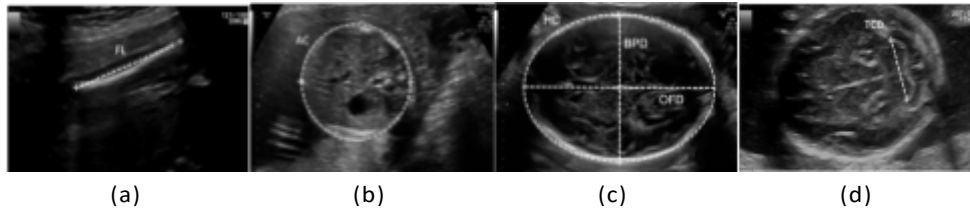


Figure 2.8: Common biometry parameters. Transverse section of: (a) femur length (FL), (b) abdominal circumference (AC), (c) head circumference (HC), biparietal diameter (BPD) and occipito-frontal diameter (OFD), (d) trans-cerebellar diameter (TCD).

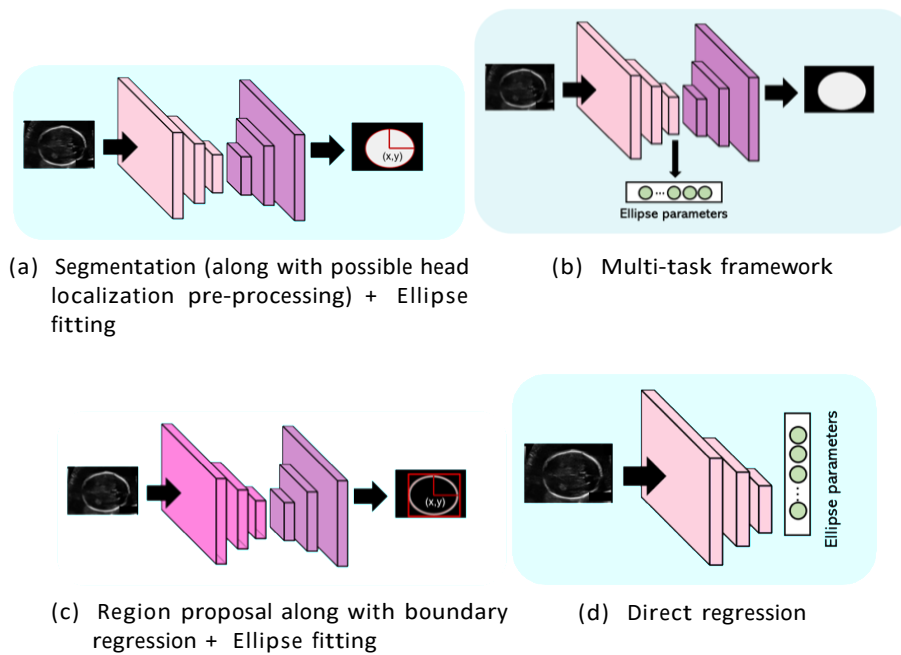


Figure 2.9: Most common deep-learning strategies to fetal HC measurement.

pregnancies) in their studies [112, 118] for in-utero surgery planning; however, clinical and pathological cases involved are not balanced.

Few papers focus the attention on multi-organ analysis: the inter-organ relations are frequently exploited by doctors when navigating and interpreting medical images [187], making the multi-organ analysis essential for a good fetus evaluation. The lack of annotated datasets, even in this field, does not allow to 1) characterize the complex inter-organ relations and 2) underline the difference between pathological and physiological images since data representative of healthy and pathological cases is critical to develop CAD systems robust to pathologies and rare anatomies.

2.5 Biometry parameter estimation

Assessing fetal size and GA, as well as detecting fetal growth abnormalities, lay the foundation of modern prenatal care. Fetal biometry assessment represents the most common medical investigation undertaken in this regard. Before 14 weeks, the GA and the fetal size are estimated by the measurement of crown-rump length (CRL), which is calculated from the top of fetus's head to the bottom of its torso. Once the CRL exceeds a fixed length (generally after 14 weeks), common measurements include HC, biparietal diameter (BPD), occipito-frontal diameter (OFD), trans-cerebellar diameter (TCD), lateral ventricles (LV), abdominal circumference (AC) and femur diaphysis length (FL). Figure 2.8 shows image samples for each of these biometries. These fetal biometries are used to assess fetal growth trajectory and ensure normal fetal development when measured at different points in time (trimesters). Additionally, cardio-thoracic ratio (CTR) and cardiac axis biometrics are measured for diagnosing CHD.

Biometry measurements are performed in a standardized manner by identifying the appropriate sonographic plane and by precisely placing calipers in the corrected position [71]. Over the years, DL algorithms that provide automated placement of the calipers to measure BPD, OFD, CRL, TCD, HC and AC have been extensively exploited to reduce operator-dependent errors and improve accuracy of fetus well-being assessment.

Most of the approaches in literature tackle the problem of caliper measurement by segmenting the area of interest at first, as shown in Fig. 2.9a.

In [153], a VGG-like CNN is used to segment fetal head and successively provide HC and BPD measurements by means of ellipse fitting. More recently, encoder-decoder CNNs have been investigated for segmentation task. Here, the majority of published papers focus on HC measurement only, relying on the HC18 challenge dataset (Sec. 2.2.2). The work in [154] proposes an active-contour model guided by external forces that are derived with a U-Net-like CNN trained to segment the fetal head. For the same goal, in [155] a V-Net is used to automatically segment the fetal skull and the HC is extrapolated. V-Net is modified to encode attention mechanism in [156]. A U-Net architecture is instead used in [157]. Similarly, in [158] a U-Net architecture is modified by adding dilated convolution layers after the last layer of the encoder and squeeze-and-excitation blocks on the skip connections. In [159] a multi-feature pyramid U-Net is proposed. An encoder-decoder network is also proposed in [160]. The network consists of dense blocks that strengthen feature propagation and channel attention in bottleneck to enhance the important features. In all these papers, HC is extrapolated by means of ellipse fitting methods.

Multi-task approaches following the scheme in Fig. 2.9b are also explored in the literature. The work in [161] proposes a multi-task CNN inspired by the structure of LinkNet, for the automatic segmentation of fetal head and the successively estimation of the HC main axes, center and angle. The authors present another approach [162]

based on a multi-scale and low complexity structure inspired by LinkNet. A multi-task approach is also exploited in [163], where head segmentation accuracy is enhanced by a feature pyramid inside a U-Net-like CNN. A regressor branch is added for predicting the HC value.

To improve segmentation accuracy, some papers incorporate a localization network in the pipeline to relieve the segmentation network from learning the position of the head. In [164], a region proposal network is used as post processing procedure once the fetal head is segmented by U-Net on polar transformed images. HC and BPD are extrapolated by an ellipse fitting method. A Mask-RCNN is used in [165] to improve segmentation accuracy by incorporating an object localisation framework. Coordinates are extrapolated by means of an ellipse fitting method.

Proposed methodologies contribute to SOTA of this field, in which a different strategy for fetal head caliper placement is considered. In [166], the issue is treated as a distance regression task, as shown in Fig. 2.9c. The proposed framework consists of a tiny-YOLO for head location and centering, and a regression CNN, which by means of distance field, accurately delineates the fetal head boundaries. HC is successively obtained using an ellipse fitting method. The work is further improved in [167] in which the framework is made end-to-end using a Mask-RCNN along with ellipse fitting for HC extrapolation. For a better comprehension of the proposed works refer to Chapter 3. The direct regression of the contours coordinates is also performed in [168]. However, CNN is jointly used with an attention refinement module along with a graph convolutional network (GCN). This approach does not requires any ellipse fitting method.

The work in [169] further extends the research in the field of HC estimation by developing two probabilistic CNN methods: Monte Carlo Dropout during inference and a probabilistic U-Net. These methods are particularly useful in the clinical practice since multiple plausible semantic segmentation of fetal heads along with HC measurements are provided to the clinicians, which can choose the best option.

In order to avoid intermediate steps, such as segmentation CNN along with ellipse fitting methods, which may be computationally expensive (both for model training and labeling), the work in [170] proposes a regression CNN to directly estimate the HC measure, as shown in Fig. 2.9d. Different CNNs are tested (including VGG16 and ResNet) as backbones. A similar approach is followed in [171] in which saliency maps are used to provide interpretation of the regression CNN results.

Besides head biometrics, a number of researchers is focusing on AC estimation. The work in [172] proposes a method for AC estimation from 2D US data. The framework performs semantic segmentation of several anatomical structures (i.e., stomach bubble, amniotic fluid, and umbilical vein) by using a custom designed CNN. AC measurement is achieved through Hough transformation, and a plane acceptance check is performed by another CNN, which verifies the presence of all structures of interest.

This method faces problems in predicting the AC precisely in the case of insufficient

amniotic fluid, which commonly occurs when observing oversized fetuses. To improve the results achieved by [172], the work in [173] proposes a method consisting of a combination of multiple CNNs.

A CNN is used to identify anatomical structures in US images (stomach bubble, amniotic fluid, and umbilical vein) and Hough transform is used to obtain an initial estimate of the AC. These data are fed to other CNNs to estimate the spine and bone positions, which are used to compute AC accurately. Then, a U-Net and a classification CNN are used to check whether the image is suitable for AC measurement.

Only one work [174] focuses on CRL measurement estimation. A U-Net is used to segment all fetus area and CRL is computed from the obtained segmentation.

Few papers in the literature focus on FL, LV biometry and CTR estimation. A first approach for automatically measuring the width of LV is proposed in [175]. This approach relies on Mask-RCNN. FL measurement is performed with SegNet and image skeletonization in [176]. The work in [177] focuses on CTR and cardiac axis estimation. The proposed CNN is a one-stage ellipse detection algorithm.

Multiple biometry estimation is addressed in [178], where head, abdomen and femur are semantically segmented using state-of-the-art CNNs. This step is followed by region fitting and scale recovery for the biometry estimation. The work in [179] proposes a multi-task framework consisting of a U-Net along with ConvLSTM to jointly localize, classify and measure HC, AC and FL in fetal US video. A multiple biometry estimation is performed also in [180] with the aim of supporting first trimester fetal assessment from a single 3-D US scan. The framework automatically extracts a slice of the whole fetus in the sagittal view and simultaneously segments the whole fetus by means of a multi-task network. Automated segmentation of the whole fetus into head, abdomen and limbs is then performed to extrapolate biometry measurements. The work in [159] is further expanded in [181], in which HC, BPD, AC and FL measurements are also considered.

The work in [182] proposes an approach that relies on unsupervised domain adaptation for HC and TCD biometry estimation from low cost US devices. The domain invariant representations for feature extraction are jointly learnt from both high-end and low cost US images and an unsupervised learning is used to calibrate the model in the low cost domain in order to produce consistent predictions. The final biometric estimation is performed fitting an ellipse to the contours (for HC measurements) and performing non-maximum suppression to find the pixels with greatest probabilities (for TCD measurements).

In [183], a fully-automated pipeline for GA estimation is proposed. The pipeline relies on BPD, HC, CRL, AC and FL estimation, which is accomplished following U-Net based structure segmentation.

An approach to automate the fetal head biometry estimation in real-time is proposed in [184]. Firstly, an AlexNet is used to classify and extract fetal head from US images; successively a U-Net is used to compute HC and BPD to estimate GA.

2.5.1 Limitations and open issues for biometry parameter estimation

Most of the papers dealing with biometry parameter estimation focus the attention on HC estimation only. Extensive research on HC parameter estimation is performed from 2D trans-ventricular planes, specifically concerning second and third trimesters screening. This may be attributed to the release of the HC18 dataset. From the virtuous example of the HC18 challenge, the research community should work to collect and release datasets for other biometries too, as very little has been done for other biometries such as CSP and CRL, instead. Table 2.7 summarizes the papers surveyed in this section.

However, the HC18 dataset does not allow studying advanced DL methodologies because of both limited number of images, anatomical structures (only fetal brain is shown) and consideration of healthy subjects only. Moreover, only one clinician participated in the annotation process, generating possible annotation bias. These issues must be addressed in order to develop a more robust research in the field. As regards parameter estimation methods, most of the papers in literature still rely on fitting method to extract object contours and successively measure biometry (as also shown in Fig. 2.9). The regression of object parameters would be a reasonable alternative since a single model to predict outputs directly from the inputs would improve the efficiency and the performance of the method. However, directly regressing single points from an input frame is highly non-linear and thus model convergence results to be more difficult to obtain [188].

Considering possible future developments, an end-to-end approach to first identify the scan plane and then calculate biometries associated to that plane could be a valuable support tool for clinicians. Only the clinical study in [183] proposes something similar, but further research is needed.

Moreover, a unified framework to biometry estimation from multiple anatomical regions could also be explored. Here, the most peculiar challenge is the huge variability in terms of shape and morphology of the anatomical structures (i.e. abdomen less contrasted with the background as opposed to other organs, femur and cerebellum that do not fit an ellipse), as well as the dimension of the organs, which varies according to the gestational trimesters.

2.6 Other tasks

Besides standard-plane detection (Sec. 2.3), anatomical structure analysis (Sec. 2.4) and biometry measurements (Sec. 2.5), researchers are working on a number of emerging tasks, which are grouped and summarized in the following subsections.

2.6.1 Gender identification

Determining fetal gender by US scans is among the main tasks addressed during the early stages of pregnancy. With the aim of canceling this information from the US

display to prevent unauthorized gender viewing, the work in [189] develops a residual network to identify frames containing gender-defining region in real-time.

2.6.2 Pose estimation

Studies about the structured description of the whole fetus in US are still rare. In [190], fetal pose is estimated by localizing 16 landmarks, including joints. The authors propose a self-supervised learning framework to fine tune a network to form visually plausible pose predictions: a pre-trained 3D U-Net predicts the heatmaps of the 16 landmarks with an intermediate fetal pose estimator. By retrieving a support set of atlases in the pose library via rigid registration, label proxies are produced to form the self-supervision. The landmark detector is tuned iteratively for on-line refinement and updated under the gradient checkpointing strategy if necessary.

2.6.3 Preterm-birth prediction

An early detection of preterm birth risks is crucial to timely intervene and preserve fetal survival. US data are typically inspected by expert doctors, which use hand-designed image features such as cervical length and anterior cervical angle. Lately, to overcome errors related to subjective method, DL was used to predict preterm birth.

In [191], a U-Net is used to segment the cervix from US images. The masks obtained are used to estimate the cervical length and the anterior cervical angle. The cervical length is measured through the centerline algorithm and evaluated performing linear regression between estimated and ground truth cervix lengths. The anterior cervical angle, instead, is estimated using an iterative algorithm, that finds centroid point of the cervix mask and splits it into two shapes: the anterior cervical angle is measured between the anterior wall and the line between the last two centroids. A similar approach is proposed in [192], which uses a U-Net with a parallel branch for classification to simultaneously segment and classify the cervix in 2D US images.

2.6.4 US simulatio

US simulation through image synthesis has a wide range of applications, including medical procedure simulation, clinical training and multimodality image registration.

Using US imaging and optical tracking data acquired on a fetus phantom, [193] simulate US images at given 3D spatial locations based on spatially-conditioned GANs, which sample anatomically accurate images conditionally on spatial position of the (real or mock) freehand US probe. Besides the US-specific difficulties, synthesising US images needs to be feasible for interactive use, which means to comply with limited computational resources. Therefore, [194] proposes a patch-based GAN that aims to improve the quality of simulated US images while keeping computational time low, via image translation of computationally low-cost images to high quality

simulation outputs. As additional input to the model, segmentation maps are provided to incorporate anatomical information potentially lost in low quality images, and also acoustic attenuation maps are given to better preserve acoustic shadows and directional artifacts information.

US image generation is explored also in [195], in which a cycle-consistent adversarial network (Cycle-GAN) is designed to simulate visually realistic third-trimester US images starting from unpaired second- and third-trimester TV US images.

2.6.5 Video summarization

Video summarization is crucial to lower the workload of clinicians when reviewing US examination. In [196], DRL is applied to video summarization to select a subset of frames to create a shorter video that contains sufficient and essential information to facilitate retrospective analysis, using an encoder-decoder CNN structure to extract relevant features from frame sequences. Video summarization is further investigated in [197], which proposes an approach for video description and clinical-workflow analysis during fetal US scans. The semi-automatic method to temporally segment the US videos into semantically meaningful segments is based on a double branch architecture in which a SonoNet is employed to extract spatial features, and an LSTM unit is used to model temporal dependency.

2.6.6 Shadow artifact removal

Acoustic shadows caused by sound-opaque occluders can potentially hide vital anatomical information in 2D US and thus can be a big burden for US analysis, ranging from anatomy segmentation to landmark detection.

An automatic shadow detection method is presented in [198]. It generates a pixel-wise shadow confidence map from weakly labelled annotations, jointly using a FCN as shadow image discriminator, a feature attribution map from a GAN and an intensity saliency map from a graph cut model. The authors improve these results in [199], developing a CNN-based, weakly supervised method for automatic confidence estimation of shadow regions in 2D US images. By learning and transferring shadow features from weakly-labelled images, continuous shadow confidence maps are directly predicted from input images. Differently from previous work, in [200] a disentanglement method is presented to disjoint anatomical from shadow features, to generalize anatomical standard plane analysis for abnormality detection in early pregnancy. The authors proposed a multi-task architecture with adversarial training, evaluated on standard-plane / shadow artifacts classification tasks.

Together with the indiscriminate mixing of image properties, e.g. artifacts and anatomy, another challenge for DL algorithms in US image analysis is the presence of different acquisition devices characteristics. The work in [201], while considering the shadow artifacts presence, explores also the cross-device adaptation problem. The authors present a non-adversarial method that evaluates mutual information

between latent features to disentangle categorical features and domain features in a semi-supervised learning framework.

2.6.7 Probe movement control

Accurate obstetric US scanning is highly operator dependent. This issue may be overcome with the use of US probe-movement guidance, especially to assist less-experienced operators. In [202], contrastive learning is used to train an end-to-end self-supervised network for visual-assisted probe movement via automated landmark retrieval. A set of landmarks is built on a virtual 3D fetal model, while during obstetric scanning a transformer is used to locate the nearest landmark through descriptor search between the current observation and the landmarks.

2.6.8 Volume reconstruction

In recent years, 3D US has become popular in prenatal care routine, commonly used to show a detailed picture of the baby and look at his face. However, 3D US could be useful for the assessment in standard planes of normal and abnormal structures, like fetal brain, spine, face, kidney and heart, to evaluate development and functionality of some organs, like heart and kidney, and to facilitate the identification of subtle fetal defects or malformations [203]. However, 3D US probe is constrained by the limited field of view and poor operability [204]. Thus, to address these problems, in [205] an approach for fetal 3D freehand US reconstruction from 2D slices is proposed. The reconstruction process is approached with a convolutional LSTM. Self-supervised learning is applied on the reconstructed volumes to perform pseudo supervision and regularize the prediction of future frames. Finally, adversarial learning is introduced to improve the representation learning of anatomical shape priors, to prevent uneven reconstructions.

2.6.9 Limitations and open issues for other tasks

Section 2.6 presented a variety of emerging approaches in the field of DL for US image analysis. Some of these approaches address interesting tasks from the clinicians' perspective, including preterm-birth prediction and fetus pose estimation. Both may provide clinicians with decision support and, if further developed, CAD. Video summarization may also be seen as a task for directly supporting clinicians, as summarization is thought to relieve clinicians mental workload and provide automatic video reporting that can be included in the health records. Other approaches focus on more auxiliary tasks, including US simulation, shadow artifact removal and volume reconstruction. These approaches are needed to increase dataset size, improve image quality or allow fast 3D volume estimation from 2D US images, but do not present an explicit clinical application. From the DL algorithm point of view, it appears that the literature is still too sparse for a proper and fair comparison. We expect,

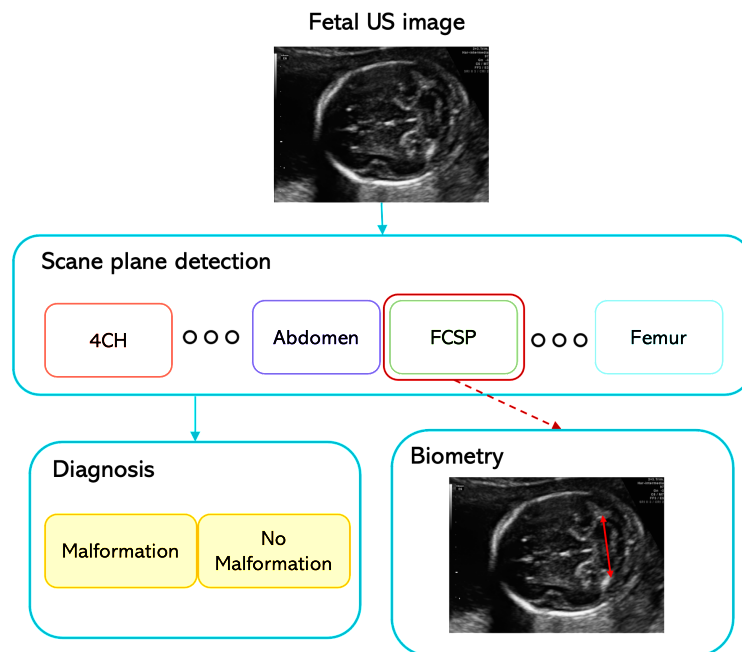


Figure 2.10: Overall workflow of a computer-assisted tool to support fetal clinicians.

however, that the research community will focus on these tasks in the next years.

2.7 Discussion and conclusions

This review analyzed a wide spectrum of the most recent DL algorithms for fetal US image analysis. Fetal US image analysis dates back to mid 1900s and a solid and rich literature today exists. Our survey started with the aim of answering the following questions:

Which are the most investigated tasks addressed using DL in the field of fetal US image analysis? According to our findings, standard plane detection (19.6%) and biometry parameter estimation (20.9%) are among the most investigated tasks (Fig 2.2a). In standard plane detection, fetal brain, abdomen and heart standard planes are the most explored since biometric measurements and identification of abnormalities are mainly performed on these planes. The recognition of the specific anatomical landmarks is the key for evaluating the quality of each scanned plane and detection algorithms or attention mechanisms are thus particularly used in this domain. As regards biometry parameter estimation, the HC is the most investigated measurement. This is mainly due to the release of HC18 challenge, which allowed a more extensive research in the domain. Segmentation (along with possible head localization pre-processing) followed by fitting methods approaches are the most exploited ones. The 49.0% of the analyzed papers dealt with anatomical structure analysis. According to Fig. 2.7, the 26,7% and the 20,0% of the papers work on

fetal heart and brain applications, respectively. The cardiac examination of fetus in literature is designed to maximize the detection of heart anomalies and genetic syndromes. Detection rate is often optimized by recognizing the main fetal heart planes (4CH, LVOT and RVOT) at first. The analysis is frequently performed on 2D US images or integrating spatio-temporal information, which improves CNN performance since the temporal dependencies present in consecutive frames are exploited.

As regards fetal brain analysis, a great part of the papers surveyed focuses on the evaluation of brain structures. The analysis is performed both with 2D and, more recently, 3D images, but structure segmentation is mainly assessed with encoder-decoder inspired architectures. GA and brain development also occupy a consistent part of fetal brain evaluation. The anatomical regions that have received marginal attention are lungs, kidney and spine, whose approaches are also less varied.

Which are the main challenges in regard to fetal examination that are currently tackled by using DL? From our survey, it emerged that DL is today able to tackle US image-related challenges, including inhomogeneities, artifacts (i.e. shadows), poor contrast and intra- and inter-clinician acquisition and measurement variability, supporting clinicians in a wide range of clinical applications. For standard plane detection, current DL algorithms are robust when processing single-center datasets, collected and annotated by 1 or 2 experts mostly from a single US machine. The algorithms are able to tackle challenges such as a high variability (size and shape) of fetuses among different GAs, as well as the presence of shared anatomical structures among different standard planes. Here, a winning strategy in the literature is to check that a predefined number of structures appears in the US probe field of view.

As regards anatomical-organ contest, DL has proven to be an essential tool to support the segmentation process of different anatomical structures, which may vary in shape and size over GAs. Boundary incompleteness is a common problem for those anatomical structures that are not particularly contrasted with respect to the background. Up to now, all the DL segmentation techniques, which are mainly supervised, outperformed other state-of-the-art methods in both performance and speed. However, these approaches may be unable to deal with boundary information loss due to a lack of prior shape information. Dilated convolutions along with residual learning are, thus, often exploited to help to increase CNN field of view and to recover information loss, respectively. CAD systems that make full use of DL have been explored, for the past few decades, as another research topic in the field, helping clinicians in their decision-making process. For biometry-parameter estimation, DL today achieves robust results on HC estimation tackling challenges such as different positions of the head in the image, varying dimensions of fetal head among the gestational trimesters and partially visible head skull. These challenges are today tackled for still images acquired from a single US machine and annotated

by a limited number of clinicians. Only one paper proposed an approach in the field invariant to significant image distribution shift between image types [182]. Besides tasks addressed in Sec. 2.3, Sec. 2.4, and Sec. 2.5, recently other tasks have been analyzed in Sec. 2.6. Emerging fields such as preterm birth prediction or fetal pose estimation are investigated to intervene and preserve fetal survival and quantify fetal movements. Techniques which improve image quality and image analysis are also exploited, such as shadow artifact removal and 3D volume reconstruction from 2D slices, which is particularly useful to provide information on the entire anatomy (typical of 3D imaging) together with a more complete field of view.

Are the commonly-used datasets sufficient enough for robust DL algorithm development and testing? Collecting and sharing annotated datasets for DL algorithm development is a well-known problem for the medical-image analysis community. Up to now, only few datasets are available in the US fetal field (Sec. 2.2.2). Besides practical issues (annotating data is a labour-intensive process), there are social and ethical concerns that have to be considered. To attenuate these issues, all these aspects should be deeply studied in collaboration with non-technical communities (e.g., lawyers, ethicists, ...). This step is crucial with a view to develop robust and generalizable algorithms.

In terms of number of images, almost all articles were evaluated with less than 1000 images and some of them even with less than 300 images. Moreover, the number of images that have been used are missing in some of the articles. In order to have a high degree of robustness, train/test size must be always clarified as well as patient number, which is an essential information to avoid that the same woman's scan is part of both training/validation and test sets. A cross-validation is strongly suggested when limited data sample is available, especially if we are facing with multi-class problems. With cross-validation strategies, all the data is used to evaluate model capability, making algorithm performance more reliable. Providing information about GA distribution is another requirement: fetus changes and grows during gestation time, thus also the evaluation of its status varies among trimesters. Routine US examinations are mainly performed during second trimester of gestation [149]; this reflects also on the analyzed papers (Table 2.5, Table 2.6, Table 2.7). The majority of them includes data acquired during 18 to 22 weeks, which serves as a baseline to evaluate fetal growth and healthy status and to perform a timely detection of major congenital anomalies. Hence, the analysis of most anatomical structures is performed on US scans of this GA and of third trimester, to evaluate and monitor compliance with standard fetal organ sizes. The only exception is the estimation of position and echostructure of the placenta, which is mainly performed during first trimester [108, 109, 111, 113, 114, 116, 117]. As regards biometry parameter estimation, the vast majority of papers deal with measurements performed during mid and third trimesters (e.g. HC, AC, FL) as opposed to first trimester ones (e.g. CRL), [174, 180], which is however of fundamental importance to estimate potential

due date. Once the fetus has passed 13 + 6 weeks, all the other measurements are performed to determine how the baby is progressing.

Which are open issues that still have to be addressed by DL in the field?

Hereafter, we discuss about the main open issues and future research directions that we have identified in the field of fetal US image analysis.

Multi-expert image annotation The lack of image annotation by multiple clinicians currently represents a major drawback. While it is undeniable that having more than one expert annotator is expensive in terms of money and time, this multi-expert annotation is crucial for the development of robust DL algorithms and for fair comparison of their performances. Currently only few papers provide a measure of the degree of intra- and inter-clinician variability. However, this parameter should be assessed as it is an indicator of the degree of image complexity. Hence, it would be advisable to always consider multiple annotations from several observers and from the same observer at different times.

This is particularly crucial when performing 1) biometry parameters estimation, since measurements are performed at pixel level requiring operator independence to be reliable; 2) CAD in anatomical structure analysis, which should not depend on only one observer/observation. With these considerations in mind, results interpretation must be carefully evaluated when manual annotation is used as the main reference for the study.

Performance evaluation The absence of a systematic evaluation workflow emerged as a critical point. Fair algorithm comparison is often hampered due to an inconsistent use of performance metrics and/ or testing datasets. A direct comparison of methods, indeed, could not be performed for all the sections under exam, since each work is validated through different measures. Moreover, despite the efforts of some researchers and International Organizations, the publicly available datasets are still too few and often released for some specific tasks (e.g., biometric measurements in a body district). The lack of available datasets in fetal anatomical structure analysis is, in fact, still to fill: addressing this issue means to increase the research in the field to a certain level, thus, providing a better evaluation of different aspects (from GA estimation up to fetal well-being assessment). DL algorithms show good performance in controlled evaluation setup, where often data from one single clinical center are available. This does not reflect the actual practice, where different US devices and probes are used, and may result in poor generalization. As for now, only few papers deal with domain generalization, proving that further research is needed.

2D data stream analysis The majority of the selected papers neglects the temporal dimension and treats each image individually. This may be considered as an issue for standard plane detection, as clinicians identify the standard plane by manoeuvring the

US probe to the desired scan plane, which results in a 2D data stream. Furthermore, processing still images may result in poor decision support to clinicians in cases in which it is important to evaluate dynamic processes, such as heart contractility analysis and fetus movement assessment.

Comprehensive analysis Developing comprehensive computational models of the fetus is particularly challenging: as the baby grows, the fetus' body is in constant transformation, with significant structural and physiological changes among trimesters leading to an important inter- and intra-organ variability. The creation of anatomically accurate models able to characterize the complexity of fetus anatomy represents, thus, one of the biggest challenges. In addition, the acquisition of quality images is the first step to adequately assess fetus well-being: an end-to-end approach able to correctly classify fetal planes and successively evaluate biometries or fetus defect has not been fully exploited yet. Among the few papers that tried to create an unified approach, the work in [88] is the most complete. However, the application is limited to only one anatomical structure (in this case the heart), and thus it does not exploit the potential of DL for distinguishing in size and shape different objects.

Computer-assisted diagnosis In CAD, DL algorithms are used to analyze images from a patient population and build a model capable of extracting information related to a certain disease and eventually provide a disease outcome prediction. Therefore, CAD may be used as a support tool in the clinical decision making process. However, data coming from pathological subjects are usually more difficult to collect compared to the ones coming from healthy subjects and building robust models performing CAD results particularly challenging. This emerged also from surveyed papers, among which only few of them included pathological subjects data, mainly related to fetal congenital heart disease and fetal brain malformations. Hence, in function of a diagnosis it is necessary to analyze fetal anatomical structures including also a wide range of pathological cases while considering also the corresponding GA information. In this perspective, researchers should work on creating balanced datasets, possibly exploring generative approaches or multicentric studies.

Semi, weak and self-supervised learning To attenuate the issue of having small annotated datasets, researchers in closer fields are proposing semi, weakly or self-supervised approaches. In the fetal US domain, only 13 out 153 papers investigate the potential of such approaches. These include [63, 65] to classify and detect planes, and [201] to explore cross-device adaptation problems. Weak supervision tasks are involved in scan plane detection [61], fetal heart representations [77], multi-organ analysis [144] and shadow detection [198, 199]. Self supervised techniques are used for scan plane detection [64], fetal pose estimation [190], probe movement estimation [202], fetal 3D reconstruction [205]. Other interesting approaches propose unsupervised learning techniques, such as in [142] for multi-organ classification and

[182] for biometry estimation in low cost US devices. Even if there is a recent attempt to exploit such techniques in the field, still too few papers make use of them, proving that this is still an unexplored field in fetal US domain.

Multimodal image analysis Integrating US with other imaging modalities may be beneficial to provide clinicians with decision support and context awareness. However, few papers in the literature addressed this topic. Multimodal image analysis may be exploited for surgical planning and guidance, for example enriching the information provided to clinicians using mixed reality, as suggested in [118].

Model efficiency The computational cost associated to training and deployment of DL models is not reported in the majority of the surveyed papers. A high computational cost hampers the efficient deployment of the DL algorithms on single-board computers for on-the-edge computation. At the same time, a high computational cost is associated with a high CO₂ consumption. Model efficiency is a critical aspect that is currently monitored by International Organizations ⁶ and performance on this regard should be considered as additional evaluation metric.

On-device DL in fetal US All the world's tech giants embrace DL to provide next-level products and services in fields such social networks, self-driving cars and finance. DL applications in medical imaging give great results in research papers, however their commercial use is not common yet. Among the DL algorithms used for commercial purposes in this field, SonoLyst is the first fully integrate AI tool in the world that allows the identification of 20 views recommended by the ISUOG mid-trimester practice guidelines for fetal sonography imaging⁷.

Use of federated learning The majority of the surveyed papers relies on single-center datasets or datasets made available through international initiatives (e.g., Grand Challenge, Sec. 2.2.2). Accessing sufficiently large and diverse datasets of fetal US images is still a significant challenge. The collaboration among clinical centers based on centrally-shared US images face privacy and ownership concerns. Federated learning is a novel paradigm for data-private multi-institutional collaborations, where model-learning leverages all available data without sharing data between institutions [206]. Despite the benefits of such a paradigm, none of the reviewed articles has implemented it. This key aspect represents a starting point for future development.

Adherence to the ethics guidelines for trustworthy AI With the rapid development of DL algorithms for fetal US image analysis, the application of ethical principles

⁶<https://cordis.europa.eu/programme/id/H2020-EU.3.5>.

⁷<https://www.intelligentultrasound.com/scannav-assist/>

and guidelines has become crucial. In 2018, the European Commission has published a white paper⁸ titled "Ethics Guidelines for Trustworthy AI" to stress on the important of respecting and promoting ethical principles in all the steps that involve the use of DL, from design to deployment. However, no attempts have been made in the reviewed papers in this directions. The Guidelines present an assessment list that offers guidance on each requirement's practical implementation. This assessment list should be filled by researchers and reported as supplementary material.

Advances in DL for fetal US analysis have demonstrated potential to support clinicians in clinical routine. From one side, DL algorithms may automatize repetitive tasks such as automatic scan plane detection and biometry estimation, allowing clinicians to save time and to better concentrate on image interpretation [207]; intra-clinician variability may also be reduced, enabling a more objective examination. From the other side, there is a growing interest in DL algorithms for complementing the current clinical protocols and enabling a more thorough examination in the future. Examples in this direction include fetal anomaly detection and 3D placenta analysis with mixed reality.

To conclude, this review introduced and discussed the most innovative and effective DL methods found in the literature for fetal US image analysis. The methods were summarized using tables reporting performance metrics, training set and test set sizes, number of annotators and related GAs. For each section, the most relevant pros and cons of each group of approaches were pointed out. We hope our review may be helpful for young researchers to get a better picture of the methods available, leading to a speed-up on the development and enhancement of methods for fetal US image analysis.

⁸<https://ec.europa.eu/futurium/en/ai-alliance-consultation.1.html>

Chapter 3

Automated measurement of fetal head circumference

This Chapter describes my personal methods to automatically measure the fetal head circumference from fetal 2D US images. As deeply discussed in 2.5, automated measurement of fetal head circumference is a crucial clinical task to assess fetus growth and its assessment is strongly dependent from clinician's expertise posing issues of both intra- and inter-operator variability and time consuming.

A large number of deep-learning approaches is addressing the problem of HC delineation through the segmentation of the whole fetal head via CNNs. Since the task is a edge-delineation problem, a different strategy based on regression CNNs was proposed. The first developed framework consists of a region-proposal CNN for head localization and centering, and a regression CNN for accurately delineate the HC.

A journal paper on this work has been published as [166]: Fiorentino, M. C., Moccia, S., Capparuccini, M., Giamberini, S., and Frontoni, E. (2021). A regression framework to head-circumference delineation from US fetal images. *Computer Methods and Programs in Biomedicine*, 198, 105771.

Moving forward with respect to the above method, an other approach which exploits an end-to-end framework based on a Mask-RCNN which though regresses HC distance fields was successively developed.

A journal paper on this work has been published as [167]: Moccia, S., Fiorentino, M.C. and Frontoni, E. (2021). "Mask-R²CNN: a distance-field regression version of Mask-RCNN for fetal-head delineation in ultrasound images." *International Journal of Computer Assisted Radiology and Surgery* 16.10 (2021): 1711-1718

3.1 Introduction

Fetal biometrics, as discussed in 2.5, have a strong diagnostic and prognostic role in the evaluation of fetal growth [208]. Among fetal biometrics, head-circumference (HC) length is measured by clinicians to estimate gestational age and fetal weight,

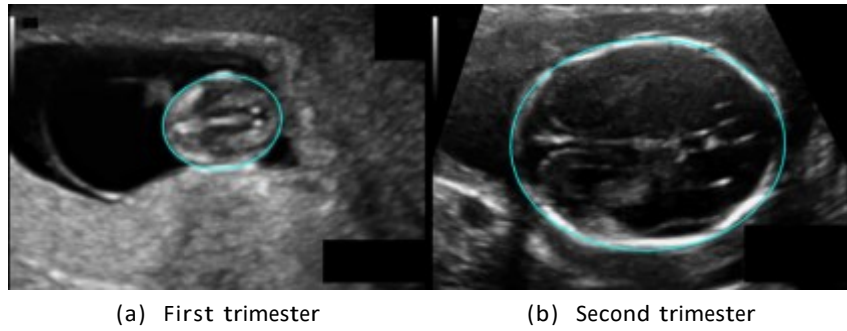


Figure 3.1: Samples of ultrasound images acquired at the (a) first and (b) second trimester. The head-circumference annotations are shown in light blue.

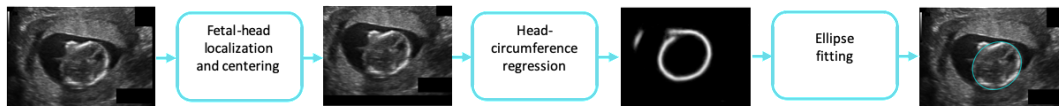


Figure 3.2: Workflow of the proposed framework for head-circumference delineation.

especially when abnormal head growth is suspected [209].

To alleviate clinicians performing HC measurements manually, different HC solutions have been proposed in literature. However, this is not a trivial task: challenges that have to be addressed include presence of shadows in the images and signal dropout and speckling, which cause fake or missing edges [40]. As a further challenge, the HC only covers a small portion of the image, which moreover varies among the trimesters (Fig. 3.1).

First approaches were mainly model-based and relied on randomized Hough transform ([210], [211]), circular shortest path [212], active contours ([213], [214]) and Chan-Vese level set [215]. More recently, machine learning has been investigated to tackle the peculiar challenges of US images. A large variety of handcrafted features has been studied. Examples include distance-[216] and intensity-based features [217], Haar descriptors ([218], [119], [41]) and textons [219].

Handcrafted-based approaches have been overcome by DL [7, 220, 221], which directly extracts features from raw data, avoiding their explicit mathematical formulation, but requires larger training datasets [1].

Research in the field of automatic HC measurement has been fostered by the *HC18 Grand Challenge*, organized in 2018, with the release of a dataset of 1334 US images, divided in 999 training and 335 test images, the latter provided without ground truth. Such dataset size, coupled with the growing availability of computational power, unlocked the potential of deep learning in the field. It is worth noting that, as the proposed researches were conducted in 2021, the examined DL SOTA approaches date back to that year. For a more up-to-date state-of-the-art version refer to 2.5.

Examples include [154], which uses CNNs to derive external forces that guide an active-contour model. This, however, may pose issues relevant to the direct

formulation of the external forces.

As a preliminary step for HC delineation, a number of researchers uses CNNs for segmenting the fetal head. The work in [161] proposes a CNN inspired by LinkNet to segment the fetal head and obtain, through adding fully connected neurons, the HC main axes, center and angle. However, the problem of directly regressing measurements may be challenging, posing issues relevant to overfitting [188]. The same authors also propose a second approach [222], where a mini-LinkNet is used. The work in [164] and [130] uses UNet-like CNNs for head segmentation, showing interesting preliminary results on small custom datasets. In [169], the *HC18 Grand Challenge* dataset is used in combination with a custom dataset of more than 2000 images to train a probabilistic UNet. As a result, multiple HC segmentation hypotheses are provided to the clinicians, which can choose the best one. Segmentation CNNs with attention mechanism are investigated in [163] and [223], showing interesting preliminary results.

A first approach to semantic segmentation has been proposed in [165], which uses Mask R-CNN for jointly localizing and segmenting the fetal head.

Researches in closer fields (e.g., [224], [225], [226]) have pointed out the benefits of modeling edge-delineation problems, such as HC delineation, as heatmap regression tasks. Following such modeling, a convolutional neural network (CNN) is trained to directly regress a distance-field from the edge of interest.

Most of these approaches mainly reduce the problem of HC delineation to that of segmenting the whole fetal head via CNNs. A different strategy could be to directly regress HC pixel position. However, this is a highly non-linear problem, which poses challenges to training convergence [188]. Regressing a distance field from the HC could attenuate this issue. This concept has recently been proved to be successful in closer fields ([224], [225], [226]). Guided by these recent considerations, in this chapter the concept of distance-field regression to accurately delineate HC from US images is investigated. In details, the chapter starts describing the first approach developed (here called 2step RCNN) [2] and continues with its final version (*MaskR²CNN*) [167]. The proposed approaches were developed and tested using the *HC18 Grand Challenge* dataset.

3.2 2step RCNN

3.2.1 Methods

Figure 3.2 shows an overview of the workflow of the proposed framework. Two consecutive CNNs for localizing and centering the fetal head (Sec. 3.2.1) and regressing the HC pixel position via distance fields (Sec. 3.2.1) are exploited. The last step (Sec. 3.2.1) consists of fitting the output of the regression CNN with an ellipse, to mimic what clinicians do in the actual clinical practice. From the fitted ellipse, the HC length is computed.

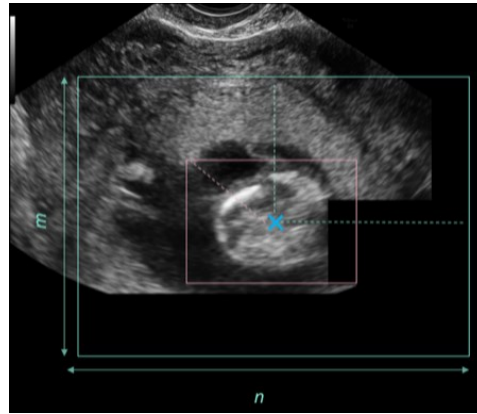


Figure 3.3: Fetal-head localization and centering: from the predicted bounding box (pink rectangle), the coordinates of the head-circumference (HC) centre are estimated. The centered image (light-blue rectangle) is obtained by centering the original image with respect to the HC center. The centered image has dimension $n \times m$ (with n and m equal to 705 and 545, respectively) to match the biggest bounding box found in the training set. Zero padding is added when the HC is too close to image borders.

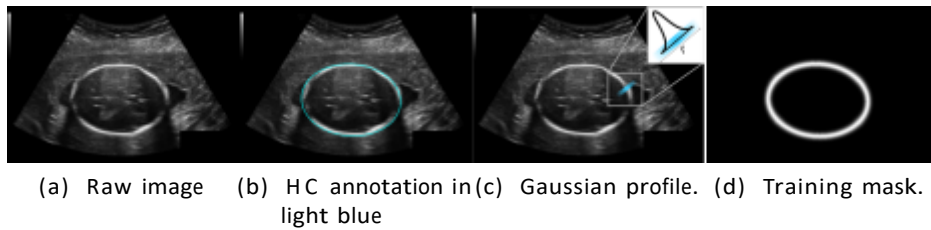


Figure 3.4: Ground-truth generation for the regression network. The raw image in (a) is annotated by drawing the ellipse representing the head circumference (HC) showed in light blue in (b). The distance field is then built following a Gaussian profile (thickness r pixels), as shown in (c), to obtain the training mask showed in (d).

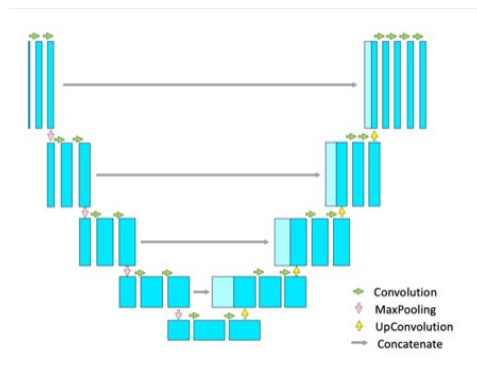


Figure 3.5: UNet architecture. The architecture includes four contraction blocks along with a bottleneck section and four expansion blocks, as described in Sec. 3.2.1.

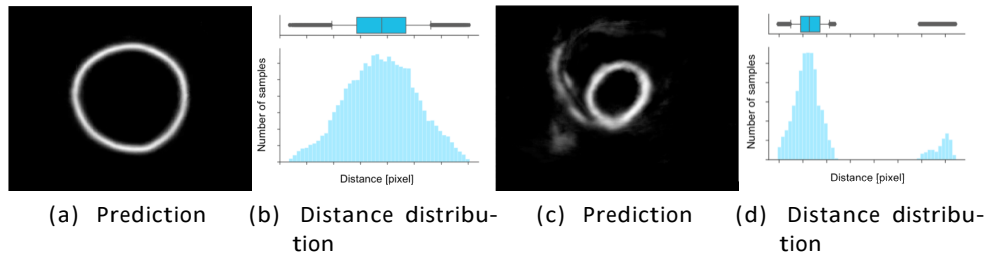


Figure 3.6: In (a) and (c), the regression output is shown for two test images. The corresponding distance distributions are shown in (b) and (d), respectively, with the corresponding boxplot and outliers on top.

Fetal-head localization and centering

As a preliminary step, inspired by [227], [164] and [165], an object-proposal localization CNN (i.e., the tiny-YOLOv2 [228]) is used for localizing and centering the fetal head. This way, the regression network is relieved of learning the position of the head.

The tiny-YOLOv2 is composed of 8 blocks, each consisting of a convolutional layer (kernel size = 3×3 , and stride = 1). Each convolution is followed by batch normalization, and the resulting output is activated with a LeakyReLU (with a constant multiplier, α , equal to 0.1 to control the slope of the activation function for negative values). Max pooling (size = 2×2) is applied after each LeakyReLU activation in the first 4 blocks, with the goal to reduce the number of parameters to learn during training.

The tiny-YOLOv2 is here trained exploiting transfer learning and fine tuning. Specifically, as initial weights, the weights obtained by training tiny-YOLOv2 on natural images are used. The weights are available online¹. For fine-tuning purpose, the weights and bias of the last 4 convolutional layers are randomly re-initialized using a normal distribution. The adaptive moment estimation (Adam) is used to minimize the loss function proposed in the original YOLOv2 paper [228]. As shown in Fig. 3.3, the bounding box provided by the tiny-YOLOv2 to center the image before feeding the regression network is used. Keeping the fetal head centered, an image region with a size ($m \times n$) equal to the that of the largest bounding box found in the training set is selected. When the region falls outside the image area (e.g., when the HC is close to image borders), zero padding is performed. This procedure is done to avoid changing the HC aspect ratio.

Head-circumference regression

The proposed strategy to train the regression CNN relies on distance fields. The rationale behind using distance field is to smooth the HC line as to facilitate the

¹<https://github.com/experiencor/keras-yolo2.git>

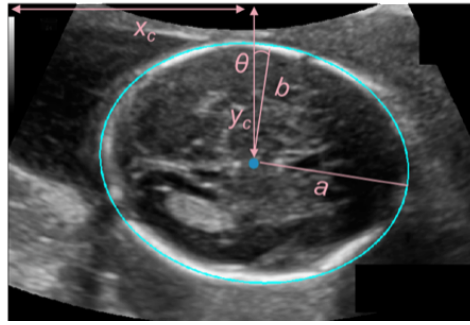


Figure 3.7: Head-circumference (HC) parameters obtained from the fitted ellipse: (x_c, y_c) = HC center coordinates, a = semi-major axis length, b = semi-minor axis length, and θ = angle of orientation.

network task with respect to directly regressing the HC line.

The HC ground-truth masks of the *HC18 Grand Challenge* dataset are 12-pixels wide. Hence, to build the distance field, the masks are skeletonized first. Then, as shown in Fig. 3.4, a region consisting of all pixels that lie in the rectangular region with thickness r pixels, centrally aligned with each of the pixel of the HC, and perpendicular to the tangent of the HC is considered.

Inspired by the work in [224], which uses a regression network to estimate surgical-robot joint position, the original UNet architecture [229] is used modifying the output layer to accomplish a regression task. Choosing UNet can be twice as beneficial: in addition to showing astounding results in regards to medical-image segmentation [230], work in the literature for fetal-head segmentation used U-shaped network, making the comparison with the proposed regression architecture straightforward. The regression architecture, shown in Fig. 3.5, consists of a contraction section, a bottleneck and expansion section. The contraction section includes 4 blocks, each with two convolution layers (kernel size = 3x3 with no padding) followed by a ReLU and a 2x2 max pooling (stride = 2). The last block adds a Dropout layer (dropout rate = 0.5) between the convolutional and max-pooling layers. The number of channels is doubled at each block to incrementally learn more complex features, starting from 64 and reaching 512 channels.

The bottleneck connects the contraction section with the expansion section using two convolutional layers (kernel size = 3x3) followed by dropout (with a rate = 0.5) and one up convolutional layer (kernel size = 2x2). The number of channel doubles from 512 to 1024.

Similarly to the contraction section, the expansion section is made of 4 blocks. Each block consists of two convolutional layers (kernel size = 3x3) followed by a ReLU activation function and an upsampling layer (kernel size = 2x2) that halves the number of feature channels. The input of each contraction section is concatenated with the correspondingly cropped feature map from the contracting section, as to recover the feature lost due to the downsampling in the contraction path. The last

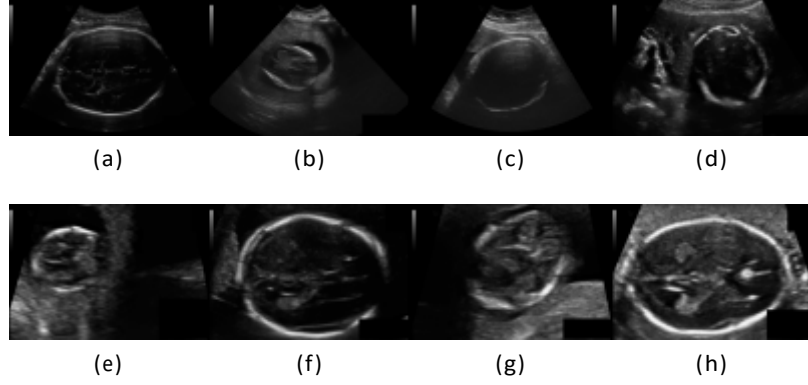


Figure 3.8: Challenges in the testing datasets include different fetus head dimensions (a, b, h) and position (d, f) in the image, partially visible head skull (c), presence of noise and US image artifacts (e, g).

block is composed of four convolutional layers, three of which are followed by a ReLU activation function. As opposed to the UNet proposed by Ronneberger [229] in which a softmax activation function is used, the last convolutional layer is activated by the hyperbolic tangent (tanh).

The regression network is trained using the stochastic gradient descent with momentum as optimizer, minimizing the Mean Squared Error (MSE) as loss function:

$$MSE(M, \hat{M}) = \frac{1}{N} \sum_{i=0}^N (M_i - \hat{M}_i)^2 \quad (3.1)$$

where M_i and $\hat{M}_i$ are the regression mask and predicted outcome from the regression network for the i -th training image, respectively, and N is number of training samples.

Ellipse fitting

Before performing ellipse fitting, the output of the regression network is thresholded. Then, the distance between each point on the thresholded output and the center of the image, which matches the HC center, resulting in a distance distribution, is computed. As shown in Fig. 3.6, the outliers of the distance distribution are identified and removed as those values outside the range $[Q_1 - O / IQR, Q_3 + O / IQR]$ [231], where Q_1 and Q_3 are the first and third quartile, respectively, IQR is the inter-quartile range, and O is a constant value that is set according to the validation set (Sec. 3.2.2).

The resulting image is fitted with an ellipse using a geometric distance based method (i.e., ElliFit) [232], which is unconstrained, non-iterative and computationally inexpensive. From the fitted ellipse, the five parameters (shown in Fig. 3.7) that univocally identify it (i.e., semi-major axis length (a), semi-minor axis length (b), angle of orientation (ϑ), and center (x_c, y_c)) are obtained.

3.2.2 Experimental setup

Dataset

The dataset used to evaluate the proposed methodology was released in the context of the *HC18 Grand Challenge*². The dataset consists of a training set of 999 and a test set of 335 US images from 551 women, acquired at the beginning of the first, second and third trimester of pregnancy. The images were collected at the Department of Obstetrics of the Radboud University Medical Center, Nijmegen, Netherlands, using both the Voluson E8 and the Voluson 730 (General Electric, Austria) [41]. All data were acquired by competent sonographers and anonymized as state in the Declaration of Helsinki. The local ethics committee (CMO Arnhem-Nijmegen) approved the collection and use of the data for research purposes.

Each image had a size of 800x540 pixels, with a pixel size ranging from 0.052 to 0.326 mm, due to sonographer adjustments to face the different size of fetuses.

For each image, the sonographer manually annotated the HC by drawing an ellipse that best fits the skull section. In this work, we kept 301 images (out of the 999 training images) as validation set.

The most peculiar challenges of the testing set are shown in Fig. 3.8. Challenges included different position of the head in the image, as well as varying dimension of the fetal head among the gestational trimesters.

Hence, the images of the first trimester showed a tiny head with skull edges not always visible. Reverberations and shadows were also be present, which resulted in poor fetal head contrast with respect to the background.

Parameter settings To train the tiny-YOLOv2, the COCO challenge annotation format was followed. Starting from the HC annotation, we generated the bounding box that bounded the HC. Prior feeding tiny-YOLOv2, the images were resized to 800x800 pixels to match the dimension of the pretrained model. The tiny-YOLOv2 was fine tuned using Adam for 100 epochs, with an initial learning rate of 0.0001 and batch size equal to 16.

Prior fine tuning, offline data augmentation (with horizontal flipping and rotation) was implemented to augment the number of images from the first semester.

The masks used for training the regression network were obtained with an r empirically set to 100 pixels and a Gaussian standard deviation of $r/2$. This allowed to fully cover the head-skull section at each HC point. Before feeding the regression network, the preprocessing explained in Sec. 3.2.1 is performed, with m and n equal to 545 and 705 pixels, respectively.

The images were zero padded (from 182x235 to 192x240 pixels) to match the dimension required by the regression network. Also in this case, padding was preferred over reshaping for preserving the fetal-head aspect ratio.

²<https://hc18.grand-challenge.org/>

Table 3.1: Metric comparison for the ablation study. The mean is reported for each metric, with standard deviation in brackets.

	Absolute difference (AD) [mm]	Difference (DF) [mm]	Dice similarity coefficient (DSC) [%]	Hausdorff difference (HD) [mm]
Shallow regression	11.11 ($\pm$ 38.25)	2.74 ($\pm$ 39.74)	93.14 ($\pm$ 12.23)	3.82 ($\pm$ 6.04)
Proposed-short	5.67 ($\pm$ 30.28)	3.62 ($\pm$ 30.60)	95.58 ($\pm$ 9.32)	2.53 ($\pm$ 4.59)
Proposed-middle	2.08 ($\pm$ 2.08)	-0.13 ($\pm$ 2.04)	97.62 ($\pm$ 4.17)	1.38 ($\pm$ 1.19)
Regression only	5.79 ($\pm$ 16.58)	5.27 ($\pm$ 16.75)	95.33 ($\pm$ 9.51)	2.82 ($\pm$ 5.47)
Proposed without excluding outliers	2.33 ($\pm$ 3.36)	1.46 ($\pm$ 4.10)	97.30 ($\pm$ 3.21)	1.51 ($\pm$ 3.68)
Proposed	1.90 ($\pm$ 1.77)	0.21 ($\pm$ 2.59)	97.76 ($\pm$ 1.32)	1.32 ($\pm$ 0.73)

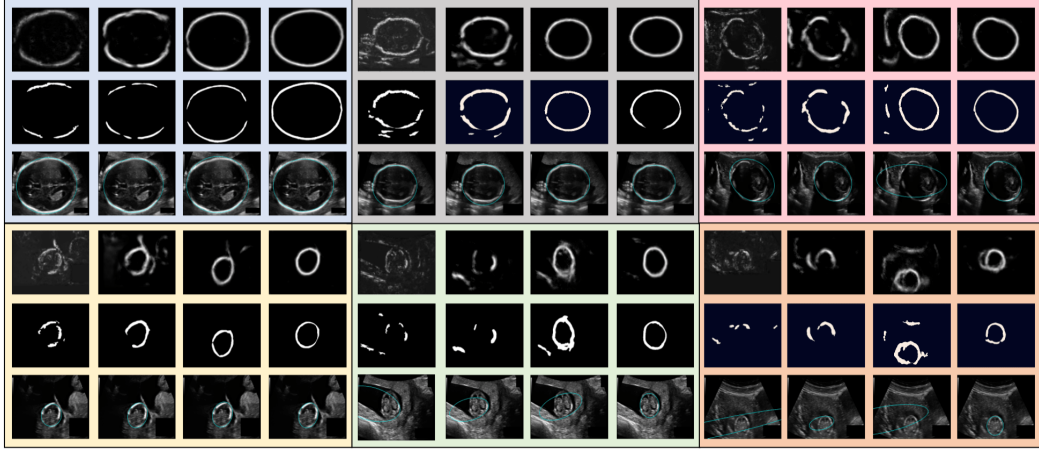


Figure 3.9: Visual samples of the ablation-study results. Each colored panel corresponds to a test image. Each panel shows the results obtained by (1st column) shallow regression, (2nd column) short proposed UNet, (3rd column) regression only and (4th column) proposed framework. Each column shows the (first row) prediction, (second row) post processed and (third-row) ellipse-fitted image.

The regression network was trained with mini-batch stochastic gradient descent, with an initial learning rate of 0.001 and a momentum of 0.98, for 1500 epochs. The value of the momentum was set according to similar work that uses regression networks [224]. Batch size was set to 32, as a trade off between memory requirements and training convergence. On-the-fly data augmentation was performed during training. Augmentation included horizontal flip and rotation in range $\pm 45^\circ$, and was randomly applied at each training iteration. The best model among epochs was selected according to the lowest mean absolute error obtained in the validation set. The threshold value was set to the 80% of the maximum value of the regression output, and the outliers for the post-processing described in Sec. 3.2.1 were identified using a O value of 0.5. The empirical threshold value of 80% and the O value of 0.5 were optimized on the validation set.

The HC length [mm] was obtained by multiplying the HC pixels for the corresponding pixel size [mm], provided by the challenge organizers. To obtain the final HC parameters, all the images were resized and recentered.

All the analyses were performed using the *Keras*³ Python library on a NVIDIA RTX 2080TI, with a Xeon e5 CPU and 128 GB RAM.

Ablation study As a first ablation study, we investigated the use of a simple regression architecture (shallow regression). The shallow regression only architecture was chosen following similar work in the literature [224].

The network was composed of 6 convolutional layers followed by batch normalization and activated with ReLU. We also tested UNet with only the 2 top layers at first (proposed-short) and then with an additional top layer (proposed-middle) (plus the bottleneck in both the architectures) to investigate the trade off between architecture depth and HC delineation performance. Both shallow regression, proposed-short and proposed-middle results are obtained including the post-processing steps presented in Sec. 3.3

To assess whether the fetal-head localization and centering helped the regression network, we compared the performance achieved by the proposed framework with that achieved by the regression network fed with raw US images.

The performance of the proposed network without excluding outliers was also tested.

For fair comparison, all approaches were investigated using the same dataset split and training setting, as well as the same computational hardware.

Performance metrics To measure the performance of the proposed method, the metrics suggested by the organizers of the *HC18 Challenge* were computed. Hence, the difference (DF) [mm] (2.9), mean absolute error (MAE) [mm] (2.7, Hausdorff distance (HD) [mm] (2.3) and Dice similarity coefficient (DSC) [%] (2.1) were computed.

3.2.3 Results

The performance comparison in terms of ADF , DF , DSC and HD of the different models proposed in the ablation study is summarized in Table 3.1. The lowest mean $AD = 1.90 (\pm 1.77)$ mm was obtained with the proposed framework, with a mean AD of $1.49 (\pm 1.32)$, $1.72 (\pm 1.50)$ and $3.25 (\pm 2.64)$ mm for images in the first, second and third trimester, respectively. The worst results were achieved with the shallow regression network, with an AD of $11.11 (\pm 38.25)$ mm. With the proposed-short and the regression-only architecture, the AD dropped to $5.67 (\pm 30.28)$ and $5.79 (\pm 16.58)$ mm, respectively. Proposed-middle achieved good results, with a mean AD of $2.08 (\pm 2.08)$. The same trend was observed for the other metrics. Without excluding the outliers from the distance distribution, as described in Sec. 3.3, the results of the proposed framework slightly dropped to (AD) $2.33 (\pm 3.36)$ mm, (DF) $1.46 (\pm 4.10)$ mm, (DSC) $97.30 (\pm 3.21)\%$, (HD) $1.51 (\pm 3.68)$ mm.

³<https://keras.io/>

Figure 3.9 shows visual samples of the results obtained from different models in the ablation study. Each colored panel refers to a test image. The first, second, third and fourth columns shows the results achieved by shallow regression, proposed-short, regression only and proposed framework, respectively. The upper, middle and lower images correspond to prediction, post-processed and ellipse-fitted images, respectively. Since no radical differences were found by visual inspection, it was decided not to include proposed-middle and Proposed without excluding outliers in the figure.

As shown in the blue panel, each of the models performed comparably when processing images with clearly-contrasted, centered HC that covers a large portion of the image. As the dimension of the HC decreased (gray panel), the shallow regression and easy UNet predictions deteriorated. The pink panel shows an image where noise and blur make the HC poorly visible with respect to the background, resulting in low performance for the shallow regression, easy UNet and regression-only architecture. The yellow and green box shows challenging images (poorly visible head skull and evident uterus edge) that compromised even further the prediction of the shallow regression, easy UNet and regression-only network. The post processing helped to attenuate the prediction inaccuracies, as shown in the yellow panel, while it was not effective when the HC texture was similar to the background tissues.

In the test image shown in the orange panel, only the proposed framework was able to delineate the head quite accurately.

3.3 Mask-R²CNN

In this section, the framework proposed moves forward with respect to [2] and hypothesize that Mask-RCNN [233], which was originally developed for semantic-segmentation tasks, can be used to provide accurate regression of HC distance fields with an end-to-end approach.

The main contribution of this strategy is a unified approach, called Mask-R²CNN (Fig. 3.10), for fetal-HC delineation in US images. The approach modifies the original Mask-RCNN by replacing the Mask-RCNN segmentation head with a new head for distance-field regression. Considering that the network regresses the distance field, it is called Mask-R²CNN because one “R” refers to the “region” proposal approach followed by the standard Mask-RCNN and the other “R” refers to the distance-field regression task. The main innovation introduced here is that this approach regresses distance fields, instead of predicting a segmentation mask as in the original implementation of Mask-RCNN. Opposite to 2step RCNN (see 3.2), the region proposal network (RPN) inherited from Mask-R²CNN avoids the need of a priori HC localization.

Furthermore, with the experiments carried out using the publicly available dataset released during the *HC18 Grand Challenge*, it is shown that the Mask-R²CNN does not produce spurious prediction. This avoids the need for a posteriori outlier removal, making the model independent from the definition of post-processing

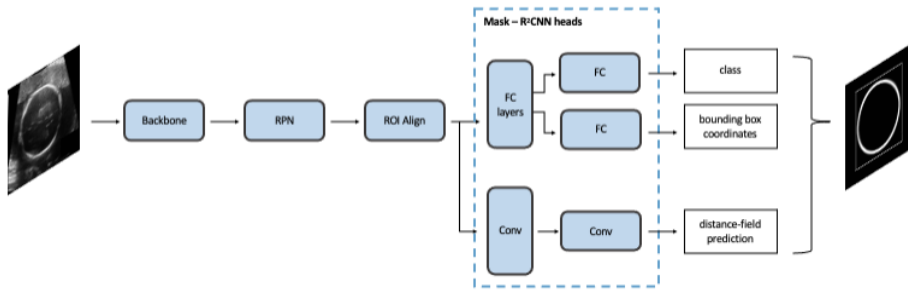


Figure 3.10: Mask-R²CNN schematic architecture. Mask-R²CNN predicts the head-circumference distance field. The relative bounding box is shown for visualization purposes. RPN: region proposal network, FC: fully connected layers, Conv: convolutional layers.

parameters. Mask-R²CNN is therefore easily generalizable to other datasets without any modification. The end-to-end approach further allows simplifying and speeding up the training process, which is an important aspect to be considered with a view to collect more US images. At the same time, having a single stage, end-to-end architecture will allow to easily embed the algorithm in US machines.

3.3.1 Method

The proposed strategy to train Mask-R²CNN relies on distance fields. As introduced in Sec. 3.1, the rationale behind using distance fields is to smooth the HC line as to facilitate the network task as opposed to directly regressing the HC line. The distance-field ground truths are the same used for 2step RCNN algorithm, described in 3.2.1.

The backbone, RPN and ROI align of Mask-R²CNN follow the standard implementation of Mask-RCNN [233].

The backbone of Mask-R²CNN is a feature pyramid network (FPN) that relies on ResNet-101. We chose this configuration as it achieved the best performance in [233]. The input US image is hence processed via a sequence of convolution and pooling. The convolutional and identity blocks follow the original implementation of ResNet [234]. The resulting feature maps (C1, C2, C3, C4, C5) are further processed by a top-down pathway with lateral connections. Convolutions in the pathway are performed with 256 1x1 filters. Up-sampling is performed with 2x2 kernels and max pooling with pool size 1x1 and strides 2. The output feature maps (P2, P3, P4, P5, P6) are processed by the RPN to generate candidates ROIs. The RPN anchors span 5 scales and 3 aspect ratios, to account for different HC size and ellipticity. Prior entering the Mask-R²CNN heads, P2, P3, P4, P5, P6 are processed by the ROI align, which resizes the candidate ROIs to guarantee that all ROIs are squared and have the same (small) spatial size $d \times d$.

Table 3.2: Mask-R²CNN distance-field regression heads. Conv = convolution, UpSamp = upsampling, dxd in the Output dimension column = spatial size of the squared feature map in output from the ROI align layer. The number of channels is reported, too.

Operator	Kernel dimension	No. filters	Output dimension
Conv	3	256	$dxd \times 256$
Conv	3	256	$dxd \times 256$
Conv	3	256	$dxd \times 2566$
Conv	3	256	$dxd \times 256$
UpSamp	2x2	-	$2d \times 2d \times 256$
Conv	2	256	$2d \times 2d \times 256$
Conv	3	256	$2d \times 2d \times 256$
Conv	3	256	$2d \times 2d \times 256$
UpSamp	2	-	$4d \times 4d \times 256$
Conv	2	256	$4d \times 4d \times 256$
Conv	3	256	$4d \times 4d \times 256$
Conv	3	256	$4d \times 4d \times 256$
UpSamp	2	-	$8d \times 8d \times 256$
Conv	2	256	$8d \times 8d \times 256$
Conv	3	256	$8d \times 8d \times 256$
Conv	3	256	$8d \times 8d \times 256$
Conv	1	1	$8d \times 8d \times 1$

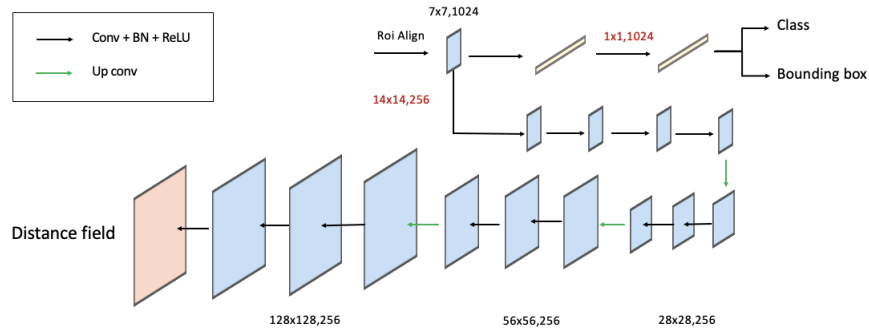


Figure 3.11: Mask-R²CNN heads. Conv: convolution, Up conv: up-sampling + convolution. Red: convolution specification (kernel dimension, number of filters), black: feature-map size.

3.3.2 Mask-R²CNN heads

Mask-R²CNN has three heads: the classification, bounding-box and distance-field regression heads, all fed with the ROI candidates from the ROI align (Fig. 3.11). In the classification and bounding-box heads, the ROI-aligned candidates are processed by two fully-connected layers with 1024 neurons. The classification head has a third fully-connected layer with 2 neurons (for fetal head and background), followed by softmax. The bounding-box head has a fully-connected layer with 4 neurons, linearly activated, which predict the anchor correction factors for the fetal-head class.

The architecture of the distance-field regression head is summarized in Table 3.2. The first four convolutions follow the implementation of the mask head of Mask RCNN. We replace the upsampling path of Mask RCNN, which originally consisted of a single transposed convolution with 256 2x2 filters with stride 2, with a sequence of up-convolutions. The upsampling path takes inspiration from the decoder path of Unet, which we used in our previous work for HC delineation [2]. Using the decoder path of Unet instead of a single transposed convolution allows us to restore the resolution of the distance field, and achieve an accurate prediction. Up-convolution is achieved with 2x2 upsampling followed by convolution with 256 2x2 filters. We set the number of up-convolutions to 3, as a trade-off between regression performance and computational effort. Each up-convolution is followed by 2 3x3 convolutions with 256 filters. All convolutions are followed by batch normalization and activation with the rectified linear unit (ReLU).

All convolutions in the Mask-R²CNN heads are performed in parallel for all ROIs using time-distributed convolution.

3.3.3 Mask-R²CNN training and ellipse fitting

Mask-R²CNN is trained using the gradient descent with momentum as optimizer and unitary batch size for memory constraint. A multi-task loss was used (L), computed on each ROI, that is defined as

$L = \alpha L_{cls} + \beta L_{box} + \gamma L_{df}$ where L_{cls} , L_{box} and L_{df} are the classification, bounding-box and distance-field regression loss, respectively and α, β, γ are constants. The L_{cls} and L_{box} are identical to those defined in Mask-RCNN. The L_{df} is the root mean square error computed between the distance-field ground truth and prediction, as in our previous work [2]. Mask-R²CNN was trained starting from the backbone pretrained on the COCO dataset, using weights publicly available online ⁴. Then transfer learning was performed by training the Mask-R²CNN heads alone, freezing the backbone. Following consideration in the literature [59], since the L_{cls} and L_{box} drop faster than L_{df} , we further trained the distance-field regression head alone, freezing the other weights of Mask-R²CNN. Finally, the full Mask-R²CNN was trained. We selected the best model among epochs as the one with the lowest L on the validation set.

Following similar approaches in the literature (e.g., [154, 161]) the output of Mask-R²CNN was thresholded prior performing ellipse fitting using a geometric distance based method (i.e., ElliFit [232]), which is unconstrained, non-iterative and computationally inexpensive. From the fitted ellipse, we derived the semi-major axis length (a), semi-minor axis length (b), angle of orientation (ϑ), and center (x_c, y_c), as required by the *HC18 Grand Challenge* organizers.

3.3.4 Experimental setup

Mask-R²CNN was developed using the data released for the *HC18 Grand Challenge* (see 3.2.2) and the same data split and data setting of 2step RCNN method. To train Mask-R²CNN, the COCO challenge annotation format was followed. Starting from the HC annotation, we generated the bounding box that bounded the HC distance field. Prior feeding Mask-R²CNN, the images were resized to 512x512 pixels, using zero padding to avoid changing the image aspect ratio. Following the original Mask-RCNN implementation, and considering the HC size in the US images, the RPN anchor scales were set to [32, 64, 128, 256, 512], with an anchor ratio of [0.5, 1, 2], where 1 means that the anchor is squared. The ROI Align output size ($d \times d$) was set to 14x14 and it was considered a total of 150 training ROIs per image, as a trade off between accuracy and memory consumption. Mask-R²CNN training was performed using gradient descent with momentum with an initial learning rate and momentum of 0.001 and 0.9, respectively. The weights of the Mask-R²CNN backbone (i.e., ResNet-101) were initialized with the pre-trained COCO weights. The α, β, γ values were set to 1 after preliminary investigation. As introduced in Sec. 3.3.3, Mask-R²CNN was trained as follows: 50 epochs for the heads (freezing

⁴https://github.com/matterport/Mask_RCNN/releases/download/v2.0/mask_rcnn_coco.h5

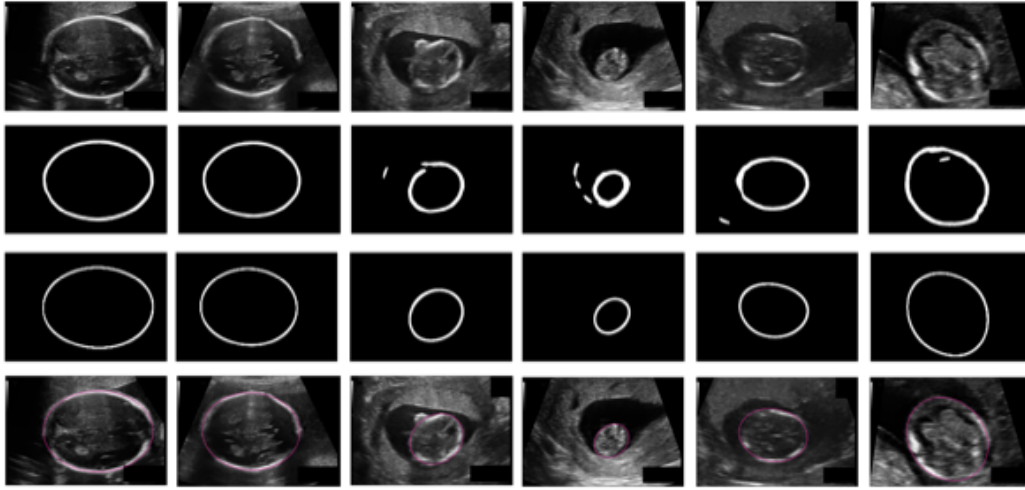


Figure 3.12: Visual samples of distance-field prediction. First row: ultrasound images, second row: prediction by [2], third row: prediction by MaskR²CNN. MaskR²CNN does not produce spurious predictions in challenging test images, avoiding the need of post-processing. The last row shows visual samples of fetal-head delineation with Mask-R²CNN.

the other layers), 50 epochs for the distance-field head (freezing the other layers) and 50 epochs for the whole network. On-the-fly data augmentation was performed using scaling, translation, rotation and shearing transformations. Thresholding prior ellipse fitting was performed using a threshold of 0.9. The HC physical length [mm] was obtained by multiplying the HC pixels for the corresponding pixel size [mm], provided by the *HC18 Grand Challenge* organizers. All the analyses were performed using *Keras*⁵ on a NVIDIA RTX 2080TI, with a Xeon e5 CPU and 128 GB RAM. The implementation of Mask-R²CNN was inspired by [235].

Following the guidelines of the *HC18 Grand Challenge*, the results were submitted to the challenge platform and computed the difference (DF) [mm], absolute difference (ADF) [mm], Hausdorff distance (HD) [mm] and Dice similarity coefficient (DSC).

3.3.5 Ablation study and comparison with the literature

The ablation study of this work is focused on the distance-field regression head. As a first study, the use of transposed convolution in the upsampling path is investigated. One (Transp1, as in the original Mask RCNN work), two (Transp2) and three (Transp3) transposed convolutions are considered. The use of upconvolution for a fair comparison with [2] are further investigated: the proposed upsampling path, which has 3 upconvolutions, with 1 (Up-conv1) and 2 upconvolutions (Up-conv2) are compared. This study allowed to find the best depth of the upsampling path. It is worth noting that the network proposed in [2] has 5 up-convolutions. Nonetheless, a

⁵<https://keras.io/>

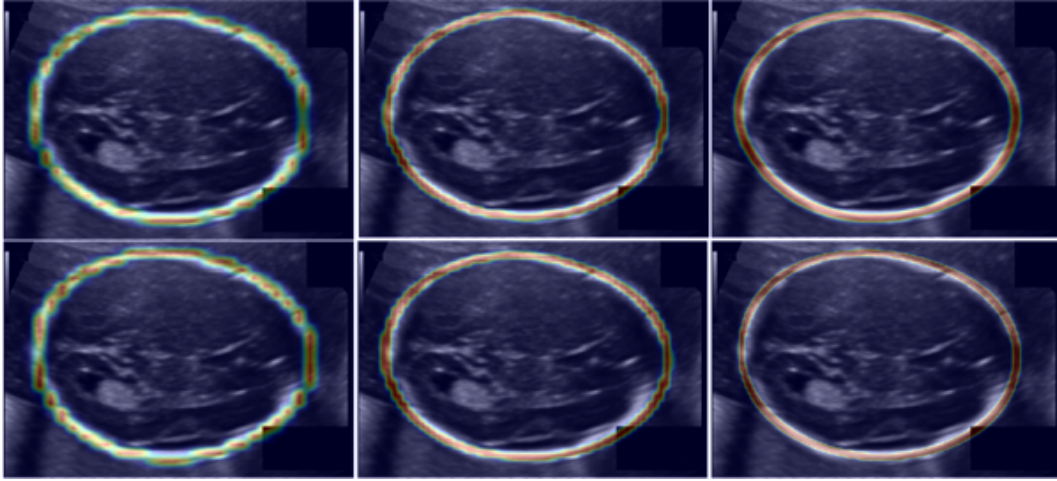


Figure 3.13: Visual samples of the predicted distance field overlapped on a test US image for each of the ablation study. A colormap is used for the predicted distance field for visualization purposes. First row (from left to right): Transp1, Transp2, Transp3. Second row (from left to right): Up-conv1, Up-conv2, Mask-R²CNN.

Table 3.3: Ablation-study results. The mean value, with standard deviation in brackets, is reported for each metric. All metrics but the Dice similarity coefficient are reported in mm.

	Absolute difference	Difference	Dice similarity coefficient	Hausdorff difference
Transp1	2.71 ($\pm$ 2.42)	-2.11 ($\pm$ 2.95)	97.44 ($\pm$ 1.10)	1.46 ($\pm$ 0.83)
Up-conv1	2.38 ($\pm$ 2.12)	-1.37 ($\pm$ 2.88)	97.33 ($\pm$ 1.28)	1.56 ($\pm$ 0.82)
Transp2	2.69 ($\pm$ 2.20)	-2.12 ($\pm$ 2.76)	97.54 ($\pm$ 1.10)	1.56 ($\pm$ 0.87)
Up-conv2	2.05 ($\pm$ 1.86)	0.15 ($\pm$ 2.77)	97.56 ($\pm$ 1.23)	1.48 ($\pm$ 0.84)
Transp3	2.08 ($\pm$ 2.05)	-0.80 ($\pm$ 2.81)	97.83 ($\pm$ 1.07)	1.32 ($\pm$ 0.77)
Mask-R²CNN	1.95 ($\pm$ 1.92)	-0.31 ($\pm$ 2.73)	97.90 ($\pm$ 1.11)	1.45 ($\pm$ 0.24)

comparison with such a number of up-convolutions was not possible due to memory constraint. As an additional ablation study, the performance of Mask-R²CNN trained without the classification head (MaskNoClass) was also evaluated. For fair comparison, the ablation study was performed using the same dataset split, training setting, and computational hardware.

The performance of Mask-R²CNN was compared against 2step RCNN [2], which is the first approach developed and the most similar to this work, and against [154], [161], [222], [163] and [165] which follow the deep-learning paradigm and use the *HC18 Grand Challenge* dataset only. We excluded the work in [223] because it uses a portion of the training set of the *HC18 Grand Challenge* for evaluation purposes. The work in [41] is also included, even if it relies on handcrafted features, because it introduced the *HC18 Grand Challenge* dataset.

3.4 Results

With the ablation study, it emerged that the best performance was achieved by Mask-R²CNN, with a mean AD , which is used for the final ranking of the *HC18 Challenge*, of 1.95 mm (standard deviation = ± 1.92) and a mean DSC of 97.90 (± 1.11). More specifically, we achieved a mean AD of 1.48 mm (± 1.39), 1.73 mm (± 1.62) and 3.62 mm (± 2.80) for images of first, second and third trimester, respectively. The testing time for one image was ≈ 0.9 s. The last row of Fig. 3.12 shows visual samples of the HC delineation obtained by Mask-R²CNN when processing challenging images in the *HC18 Challenge* test set. HC delineation is shown for fetal heads of varying size and position in the image. The Transp3 (2.08 mm (± 2.05)) and Up-conv2 (2.05 mm (± 1.86)) achieved closer AD than Mask-R²CNN. The worst performance was achieved by Transp1 (2.71 mm (± 2.42)) and Transp2 (2.69 mm (± 2.20)). Visual samples of the distance-field prediction for the ablation study are shown in Fig. 3.13. The distance fields obtained with Transp3 and Mask-R²CNN granted the highest resolution. MaskNoClass achieved a mean DSC and AD of 82.31 and 4.14, respectively, achieving the lowest performance in the ablation study.

For the sake of brevity, the performance relevant to the comparison with the literature are reported in terms of AD . Mask-R²CNN outperformed [41] (mean AD = 2.80 mm (standard deviation = ± 3.30)), [161] (2.12 mm (± 1.87)), [222] (2.22 mm (not available)), [154] (2.45 mm (± 2.55)), and [165] (2.33 mm (± 2.21)). Our Mask-R²CNN performed slightly worse than [163], which is the best-performing method published so far, with a difference in AD estimation of 0.14 mm. Mask-R²CNN also outperformed our previous work [2] when excluding its post processing (2.33 (± 3.36)). Nonetheless, also when including the post processing of [2], Mask-R²CNN had higher DSC (97.90 (± 1.11) vs 97.76 (± 1.32)) and close AD (1.95 mm (± 1.92) vs 1.90 mm (± 1.77)). Visual results of the distance-field regression output for Mask-R²CNN and [2] are shown in Fig. 3.12. The method proposed in [2] produced spurious predictions both inside and outside the HC area requiring the post processing step, while Mask-R²CNN produced accurate distance-field estimation.

3.5 Discussion

In this work, an end-to-end deep-learning approach to accurately delineate HC in fetal ultrasound images was presented. The approach is based on distance field, moving our previous approach [2] forward: a modified MaskR-CNN, called Mask-R²CNN, was built by replacing the segmentation head with a new head for distance-field regression. To tackle challenges related to low resolution of the predicted distance field, which would hamper accurate HC delineation, the original upsampling of Mask RCNN was replaced taking inspiration from the upsampling path of UNet. This end-to-end approach paves the way for embedding Mask-R²CNN directly in US devices with short inference time, not requiring any pre- or post-processing.

Mask-R²CNN was developed using the *HC18 Grand Challenge* dataset. The dataset presented multiple challenges, including poorly visible HC, different HC location and dimension as well as presence of reverberations, speckles and shadows. Despite such challenges, the results of Mask-R²CNN were encouraging with a mean AD of 1.95 mm (standard deviation ± 1.92), proving it to be a perfect competitor compared to other HC delineation framework in the literature. Mask-R²CNN performed slightly worse on third trimester images (mean AD of 3.62 mm (± 2.80 mm)) as opposed to mean AD of 1.48 mm (± 1.39 mm) obtained in the first trimester). This may be due to the fact that image pixel dimension is higher on third-trimester images compared to those belonging to other trimesters. This was also found in [2].

With the ablation study, the higher the number of up-convolutions or transposed convolutions, the more accurate the Gaussian profile regression, as shown in Fig. 3.13. The worst results in terms of AD were in fact achieved with Transp1 (i.e., using a single transposed convolution as in the original Mask RCNN) with a mean AD of 2.71 mm (± 2.42 mm), which was followed by Transp2 and Up-conv1 with a mean AD of 2.69 mm (± 2.20 mm) and 2.38 mm (± 2.12 mm), respectively.

Up-convolutions seemed to performed slightly better compared to transposed convolutions. Hence, from our experiments up-convolution guaranteed a better distance-field output resolution than transposed convolution, allowing to have a more accurate delineation of the fetal head contours. This may not be fully appreciated considering the DSC , as it was computed from the full-head segmentation masks. Considering the AD , which is the elected metric for assessing the best method by the *HC18 Grand Challenge* organizers, it was always lower for up-convolution than for transposed convolution. This can be also appreciated from a qualitative point of view by the visual samples shown in Fig. 3.13. MaskNoClass achieved the lowest results in the ablation study. This may be probably explained considering that the classification loss of our Mask-R²CNN had a regularization effect during training. This is in line with current considerations in the literature about multi-task learning [59, 236].

Mask-R²CNN was one of the best HC delineator among state-of-the-art methods tested on the same dataset. The lower performance of [41] and [154] may be explained considering that deep learning is more robust to US challenges than handcrafted-based and model-based strategies. As regard lower performance of [161], directly regressing the HC parameters, without going through an intermediate step, could be challenging for the architecture. The proposed approach also outperformed [222] and [165], in which the HC delineation was performed as a segmentation problem. Regressing a distance-field by Mask-RCNN may be therefore a satisfactory way for HC length computation. Our approach performed slightly worse compared to [59], with a difference of 0.14 mm in AD estimation which may be probably attributed to the use of attention mechanism. However, a difference of 0.14 mm can be considered negligible compared to the whole size of the head (mean HC = 174.38 mm among the images of the training set). Mask-R²CNN also outperformed our previous work [2]

when its post processing was not considered. The presence of a RPN allowed to obtain accurate predictions (especially in those images in which the uterus is particularly evident (see columns 3-6 in Fig. 3.12)), avoiding to rely on a post processing method to discard outliers. Nonetheless, even when the post processing was considered, the proposed framework reached a close AD , with a difference of 0.05 mm. Moreover, since the up-sampling path of [2] included one more up convolution compared to our Mask-R²CNN, adding one more up-convolution is supposed to further improve the distance-field prediction. This was not tested in this work due to memory constraints. Even though accurate and fully embeddable, the main limitation of the work is its focus on only HC biometry in "not abnormal" subjects. As also stated in 2, this extrinsic lack of generalizability over other fetal organs and pathological cases lies on the deficiency of acquiring different fetal US samples. Data collection is in fact a trivial task which may be impractical for different reasons: i) lack of experience when dealing with annotation ii) lack of data acquisition laboratory. The publicly release of fetal US datasets represents one of the main challenges researchers have to focus on. Generative-based methods have recently been exploited as a data augmentation technique to attenuate the issue of the paucity of data in the medical image domain [237]. Exploiting synthetic augmentation techniques as proposed in [238] may be a possible alternative.

To directly delineate the HC, hence avoiding to rely on ellipse fitting, semantic edge localisation could be also investigated.

3.6 Conclusions

In this work, we showed that exploiting regression CNNs for HC delineation from US images may be a valuable solution to automatically generate HC measurements. We achieved an $AD = 1.90 (\pm 1.76)$ mm for 2step CNN, which however has the drawback of not being an end-to-end approach. Mask-R²CNN is able to tackle the challenges of HC delineation in US images, achieving an AD of 1.95 mm, without any manual intervention nor pre or post-processing. We moved forward from our previous work [2] presenting an end-to-end architecture that can be easily embedded in US machines and used also in other clinical fields. It is to be hoped that Mask-R²CNN could be translated in the clinical practice to offer true support to clinicians for HC measurement.

Chapter 4

On the use of contrastive learning for standard-plane classification

This chapter studies the feasibility of contrastive learning approach for fetal standard plane detection (a fetal application already presented in Chapter 2). Identifying standard planes during fetal ultrasound acquisition is the first step for fetal biometry measurement and organ evaluation. Even if an extensive literature on this topic already exists (Section 2.3), the problem of relying on time-expensive data annotation has not been fully solved in this domain. In other fields of medical-image analysis [239], contrastive representation learning based on instance discrimination tasks [240] has gained much attention to incorporate unlabelled data in the training phase and improve classification performance despite the limited amount of labeled data.

This chapter aims at studying the effectiveness of these approaches in a field in which data is not readily available, nor are annotations like fetal SPD. In particular, semi supervised approaches which integrate self-contrastive learning paradigm are investigated.

The method relies on ResNet50 for plane classification and SimCLR for contrastive learning. The classification performance is assessed when (i) pretraining ResNet50 backbone on ImageNet (without SimCLR), (ii) training ResNet50 backbone from scratch using SimCLR, (iii) like ii, but starting from ImageNet weights, (iv) training an overall framework for contrastive learning and classification in a dual-task fashion. We investigate how the classification performance changes when considering different percentages of training patients, and datasets with different degrees of inter-class variability. The analyses are performed on the most recent publicly-available dataset in the field.

A conference paper on this work has been submitted in IPCAI 2023 conference as: Migliorelli, G. and Fiorentino, M.C., Villani, F.P., Di Cosmo, M., Frontoni, E., and Moccia, S. (2023) "On the use of contrastive learning for standard-plane classification in fetal ultrasound imaging" and it is currently under review.

4.1 Introduction

Ultrasound (US) imaging is widely used for fetus monitoring due to its low cost and safety. It allows clinicians to assess the development of all fetal organs and maternal structures, and perform early screening of fetus malformation.

According to the guidelines of the international society of ultrasound in obstetrics and gynecology (ISUOG) [71], identifying fetal standard planes is crucial for fetus assessment reproducibility. Today, this task is performed by clinicians, who move the probe across mother's belly, searching for specific anatomical landmarks. Typical acquired planes include maternal cervix, FFESP, thorax and FBSP. FBSP include TV, TT and TC planes (Fig. 4.1). From the clinicians' side, identifying standard planes is challenging due to the high inter-subject variability and the low inter-class variability, especially when considering brain planes. Even though really similar, each fetal brain standard plane is in fact characterized by the presence of different anatomical landmarks. All these characteristics can be visualized at the time of the head measurements or brain anatomical integrity [149]. TT and TV allow the assessment of the cerebral hemisphere regions (thalamus and atrium, respectively). TC plain permits the evaluation of the cerebellum (see Figures 4.1). Other challenges include the presence of common US artefacts that influence image quality (e.g. reverberations and speckle noise) [241, 242].

To assist clinicians in identifying standard planes, deep learning has been extensively used, as pointed out by the review paper on fetal US image analysis [14] (Chapter 2).

Tackling the lack of large annotated datasets has been identified as one of the main obstacle to the translation of developed algorithms to the actual clinical practice. As for now, in fact, only one dataset [39] is currently available in the field.

In other fields of medical-image analysis (e.g., [239, 243, 244]), contrastive representation learning based on instance discrimination tasks [240] has gained much attention to incorporate unlabeled data in the training phase and improve classifi-

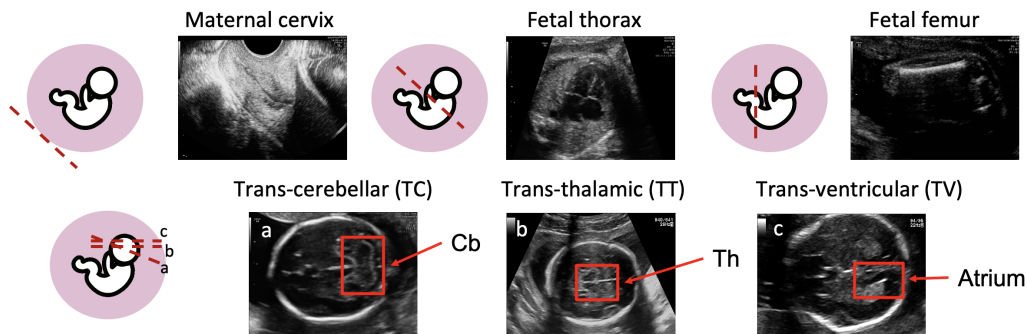


Figure 4.1: Main fetal standard planes are shown as well as the main different anatomical landmarks among brain standard planes. Cb = cerebellum, Th = thalami.

cation performance despite the limited amount of labelled data. However, in the field of fetal US image analysis, there is still a number of open questions relative to contrastive learning such as:

1. *To which extent using contrastive learning improves detection performance over standard transfer learning from large datasets of natural images?*
2. *Should inter-class variability be taken into account to decide whether to use contrastive learning?*
3. *May contrastive learning be beneficial even though when working with a small dataset?*

Thus, guided by these questions, this work studies contrastive learning for semi-supervised classification conducting a fair comparison among i) self-supervised pre-training, ii) supervised pre-training on ImageNet iii) self-supervised pre-training starting from ImageNet weights and iv) training an overall framework which uses contrastive learning and classification in a dual-task fashion. We show that the impact of self-supervision on the classification performance strongly depends on the inter-class variability of the experimental dataset and percentage of unlabelled data. Our results provide suggestions to medical image researchers on when to rely on self-supervision as opposed to pretraining on natural images. Our all experiments are performed on the only currently publicly available dataset in the field (released in 2021).

4.2 Related work

Today, an extensive literature on fetal standard plane classification exists [14] and some papers exploit unlabelled data to train CNNs in conjunction with a small number of labelled images/videos. Pretext tasks used to generate useful feature representations are the most investigated strategies to take advantage of unlabelled data. The work in [245] proposes an approach in which a model is designed to correct the order of a reshuffled video clip, along with the prediction of the geometric transformation applied to the video clip. The learned meaningful and strong representations are then transferred to the downstream task to classify 13 standard planes. In [64] two small patches are extracted from a 2D US image and randomly swapped. Model has to restore the altered image back. The weights obtained from this pretext task are successively used to perform classification of several standard planes. Even though successful, the designed pretext tasks may limit the generality of the learned representations[246]. A different approach is proposed in [63] by enforcing similarity between data coming from the same labelled class. Then, a semantic transfer objective is used to transfer this information to unlabelled datapoints by entropy minimization. A consistency loss is used in [65] to classify 13 standard

planes. In particular, predictions from unlabelled data, are used as labels for the same augmented images. This allows to take into account class imbalance, background class and different levels of inter-class similarity.

Self-supervised contrastive learning has recently become viable in natural images domain being able to learn latent representation of images without any label. These approaches have also been used in US [247], where, to avoid similarity conflict, two samples from the same video have been taken as a positive pair and two samples from different videos are regarded as a negative pair. However, even if recently used in US domain, no one has already deeply studied when contrastive learning is beneficial in fetal standard plane discerning.

4.3 Method

In this section, the impact of contrastive learning for standard-plane classification is evaluated by comparing different training strategies. In all our experiments, ResNet50 is used as standard-plane classifier. All the experiments are performed on the only publicly-available dataset for the task of fetal US standard plane detection¹ [39].

Dataset The dataset used in this research is one of the available ones in the field [39](Section 2.2.2). It includes routinely acquired maternal-fetal screening 2D US images from 6 classes: abdomen, brain, femur, thorax, maternal cervix and other. Images are manually labeled by an expert fetal clinician. In our experiments, we consider a portion (Z-multiplane) of the original dataset that includes maternal cervix, brain, thorax and femur images, while we exclude the miscellaneous Other class as well as the abdominal class, due to its low number of samples. Z-multiplane grants high inter-class variability, as shown in Fig. 4.1. The number of train and test images in Z-multiplane is 4175 (852 patients) and 3301 (894 patients), respectively. The split is provided by [39]. The subset of brain images (Z-brain) in Z-multiplane is further considered as a second experimental dataset. Images in Z-brain are labeled as TC, TT and TV and show low inter-class variability (Fig. 4.1). In this case, a total number of 1543 images (524 patients) and 1406 (536 patients) is available for training and testing, respectively. Also in this case, the split provided by [39] is used.

Table 4.1 and table 4.2 show the training set size and the testing set size for each class in Z-multiplane and Z-brain, respectively. For both Z-multiplane and Z-brain, all images are resized to 192x288 pixels. For validation, we select the images from the 20% of training patients.

Contrastive learning Our approach relies on SimCLR for contrastive visual representation learning [246].

Following SimCLR implementation, in a batch of N US images, each image x_i with $i \in [0, N - 1]$ is augmented twice into $\hat{x}_{2i}$ and $\hat{x}_{2i+1}$, through two randomly

¹<https://zenodo.org/record/3904280>

Table 4.1: Number of training images and patients for each class.

Percentage	Multiplane			Brain plane		
	Plane	Images	Patients	Plane	Images	Patients
10%	Fetal brain	169	55	T C	48	34
	Fetal thorax	103	48	T T	87	44
	Maternal cervix	92	51	T V	29	21
	Fetal femur	57	38			
25%	Fetal brain	394	133	T C	107	85
	Fetal thorax	248	111	T T	216	111
	Maternal cervix	227	128	T V	73	57
	Fetal femur	57	38			
50%	Fetal brain	765	261	T C	187	154
	Fetal thorax	483	219	T T	440	228
	Maternal cervix	452	246	T V	144	116
	Fetal femur	224	175			
75%	Fetal brain	1212	401	T C	274	224
	Fetal thorax	780	335	T T	665	349
	Maternal cervix	718	373	T V	216	172
	Fetal femur	398	284			
99%	Fetal brain	1608	531	T C	371	304
	Fetal thorax	1049	438	T T	867	455
	Maternal cervix	976	495	T V	289	227
	Fetal femur	511	367			

Table 4.2: Number of images and patients for each class in the test set

Plane	Multiplane		Plane	Brain plane	
	Images	Patients		Images	Patients
Fetal brain	1472	546	T C	339	267
Fetal thorax	660	312	T T	765	451
Maternal cervix	645	418	T V	302	215
Fetal femur	524	384			

Table 4.3: Data augmentation used during contrastive learning (strong augmentation) and classification (weak augmentation). Ranges for each transformation are reported along with the augmentation probability.

Transformation	Weak augmentation	Strong augmentation	Probability
Brightness	[-0.2, 0.3]	[-0.4, 0.5]	1.0
Contrast	[0.9, 1.0]	[0.7, 1.0]	1.0
Horizontal flip	-	-	0.5
Crop (%)	[0.9, 1.0]	[0.7, 1.0]	1.0
Rotation (deg)	[-12.0, 12.0]	[-15.0, 15.0]	1.0
Translation (pixel)	[-10, 10]	[-10, 10]	1.0

parameterized strong augmentations $T1$ and $T2$. $T1$ and $T2$ can be random color augmentations, crop with resize, rotations, translations and random flips, as summarized in Table 4.3. $\hat{x}_{2i}$ and $\hat{x}_{2i+1}$ are named views and provide a positive pair. The views obtained from all the other j images with $j \in [0, N-1]$ and $j \neq i$ are the negative ones.

The $\hat{x}_{2i}$ and $\hat{x}_{2i+1}$ are fed into a Siamese network ($f(\bullet)$), which extracts representation vectors $h_{2i} = f(\hat{x}_{2i})$ and $h_{2i+1} = f(\hat{x}_{2i+1})$ where h is the output after global average pooling. These vectors are fed into a ReLu-activated Multi Layer Perceptron with one hidden layer, yielding two embeddings (z_{2i}, z_{2i+1}).

The normalized temperature scaled cross-entropy (NT-Xent) loss (L_{contr}) is used to minimize the difference between any embedding and its positive sample while maximizing the difference between the embedding and all negative samples:

$$L_{contr}(i) = -\log \frac{\exp(\text{sim}(z_{2i}, z_{2i+1})/\tau)}{\sum_{k=0}^{N-1} \mathbf{1}_{[k \neq 2i]} \exp(\text{sim}(z_{2i}, z_k)/\tau)} \quad (4.1)$$

where $\text{sim}(\mathbf{u}, \mathbf{v}) = \frac{\mathbf{u}^T \mathbf{v}}{\|\mathbf{u}\| \|\mathbf{v}\|}$, $\mathbf{1}_{[k \neq 2i]} \in \{0,1\}$ is an indicator function equal to 1 for $k \neq 2i$ and τ is the temperature parameter.

Training strategies The following experiments using both the Z-brain and Z-multiplane datasets are performed:

- Experiment 1 (E1 - Z-brain, Z-multiplane): pretrained ResNet50 backbone using ImageNet weights, to assess the benefit of transfer learning using a large natural-image dataset. In this configuration, only the cross-entropy loss (L_{CE}) is minimized using ADAM optimizer. Weak data augmentation is used (Table 4.3), as suggested in [246]. L_{CE} is minimized using a learning rate (LR) equal to 0.001 freezing the ResNet50 backbone for 50 epochs. For additional 150 epochs, training is performed with LR = 0.0001 unfreezing the backbone.

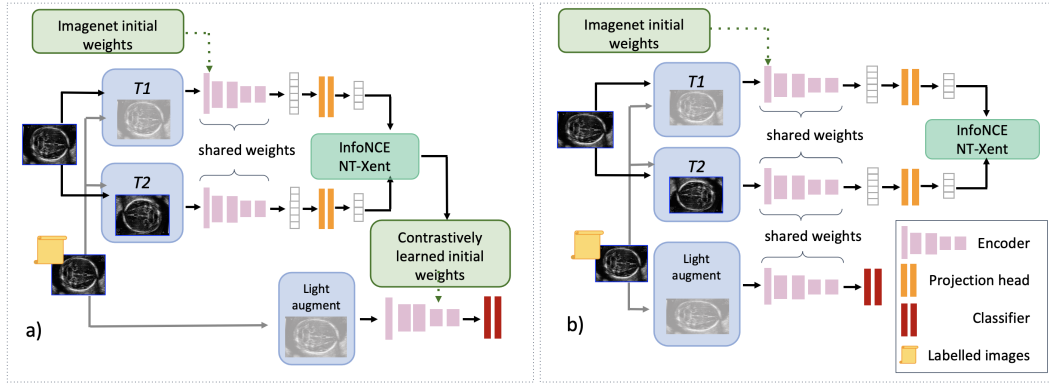


Figure 4.2: Visual representation of E3 (left) and E4 (right) experiments described in Sec. 4.3. E2 experiment follows what is currently done in E3 without ImageNet weights initialization.

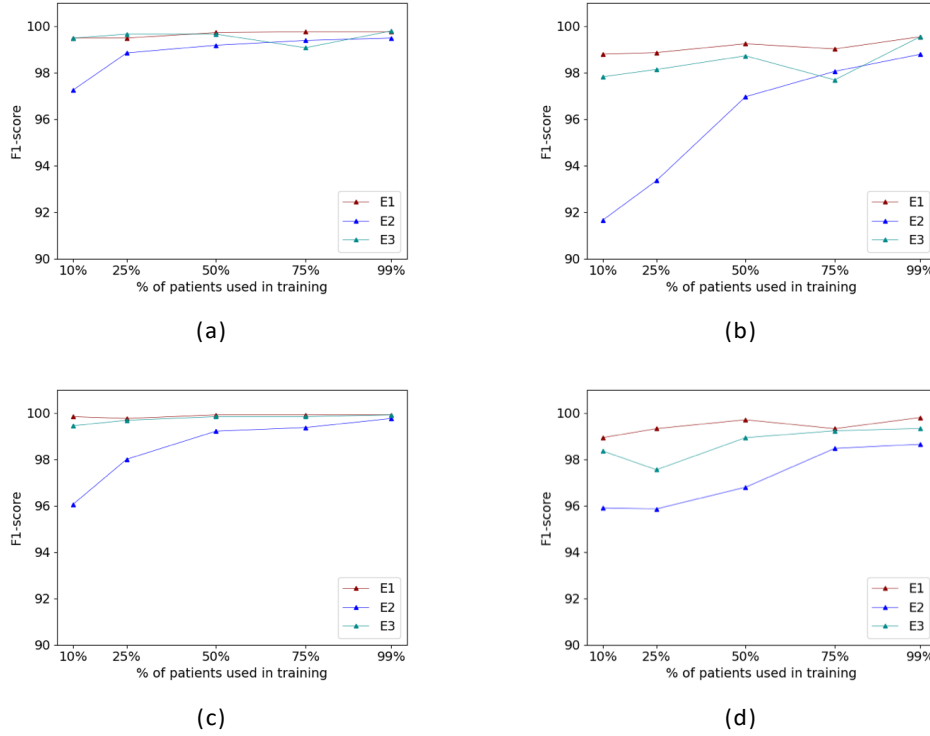


Figure 4.3: F1-score obtained with E1, E2 and E3 - Z-multiplane considering different percentage of training data. Results are shown for (a) fetal brain, (b) fetal thorax, (c) maternal cervix and (d) fetal femur.

- Experiment 2 (E2 - Z-brain, Z-multiplane): ResNet50 backbone trained from scratch using simCLR with L_{contr} , to evaluate the effectiveness of initializing backbone weights using contrastive learning. To accomplish the classification task, ResNet50 is fine-tuning using L_{CE} from weights initialized with simCLR. The L_{contr} is minimized for 150 epochs with LR equal to 0.001, while training

L_{CE} procedure follows E1.

- Experiment 3 (E3 - Z-brain, Z-multiplane): E3 replicates E2 but training from scratch is replaced with pretraining using ImageNet weights, to assess whether combining transfer learning and contrastive learning may be beneficial (see Figure 4.2a).
- Experiment 4 (E4 - Z-brain, Z-multiplane): End-to-end dual-task training including a pretext-task based on contrastive learning and a classifier, both sharing the same backbone (see Figure 4.2b) whose initial weights are those from ImageNet. Under a transfer learning scenario and in order not to waste the advantage brought by ImageNet, training is split into two consecutive steps. During the first step, the backbone is frozen and 50 epochs are consumed with a $L R = 0.001$. During the second step, the backbone is unfrozen and fine-tuning is run for 150 epochs with a $L R = 0.0001$ in order to not vanish the benefits so far gained. Batch size is set to 32 as all the other experiments.

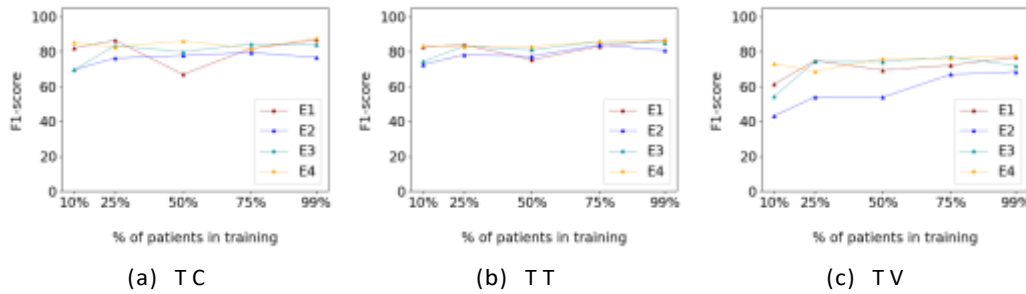


Figure 4.4: F1-score obtained for E1, E2, E3 and E4 for Z-brain.

Each experiment is performed using images from different amounts of patients and each model among epochs is selected according to the lowest loss obtained in the validation set. Percentages of patients in the training set are 10% (161 images), 25% (361 images), 50% (771 images), 75% (1155 images), 99% (1527 images). All experiments are run on two NVIDIA RTX A6000 GPUs and two NVIDIA GeForce RTX 2080 Ti. The code was implemented using Tensorflow 2.6.0.

Classification performance is assessed quantitatively by means of $F1$ to deal with class imbalance. ANOVA test is performed to evaluate the presence of significant differences among the four configurations.

The classification performance is also evaluated qualitatively using class activation mapping (CAM).

4.4 Results and discussion

Figure 4.3 reports the $F1$ obtained with E1, E2 and E3 with Z-multiplane. All experiments showed comparable performance for each of the 4 standard planes, even

4.4 Results and discussion

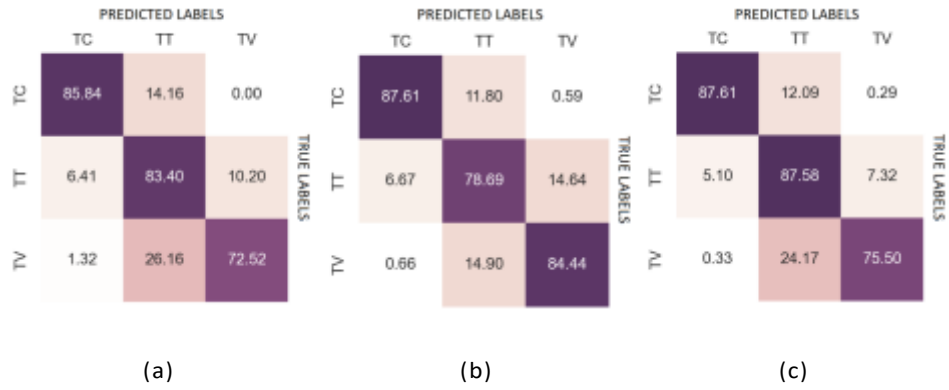


Figure 4.5: Confusion matrices for E4 and Z-brain considering the (a) 10%, (b) 50% and (c) 99% of training patients.

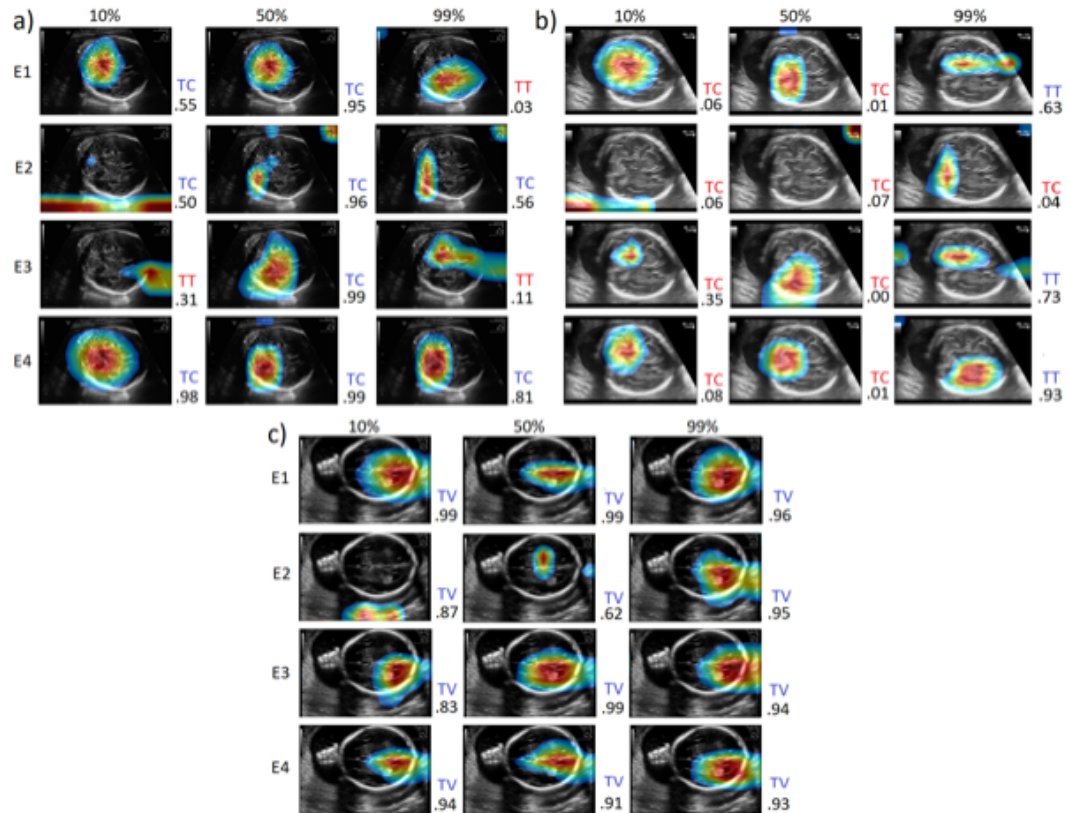


Figure 4.6: Visual samples of CAMs obtained from E1, E2, E3, E4 experiments for different percentage of annotated patients. Block a), b) and c) show results for an image of class TC, TT and TV, respectively. Numeric values provide the probability gained by the true class (blue labels), whereas red labels identify the predicted class whenever it differs from the ground truth.

when considering different percentages of training patients. A plateau in the $F1$

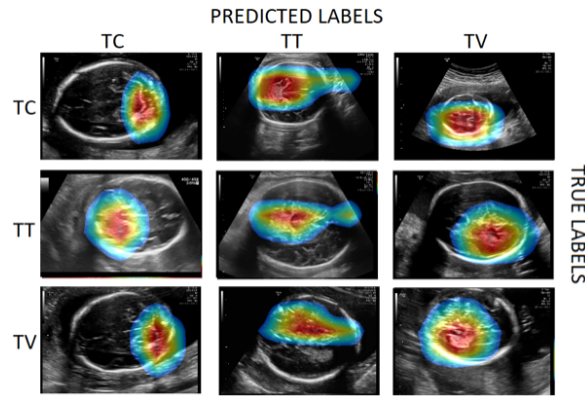


Figure 4.7: Visual confusion matrix with sample images along with the corresponding CAM.

(higher than 95%) was achieved once the 15% of the training patients was used. Considering this plateau, we did not report E4 results. Such results may be explained considering the high inter-class variability of Z-multiplane. In fact, the recognition of really different images (e.g. presence of rounded structures for brain versus more linear shapes in the case of fetal femur as shown in Figure ??) does not require complex training strategies.

Figure 4.4 shows the result of E1, E2, E3 and E4 when processing Z-brain. As expected considering the lower inter-class variability of Z-brain as opposed to Z-multiclass, an overall lower average $F1$ was reached, as also reported in [39]. TT and TC planes had the highest $F1$ in all experiments (see Figure 4.4 a-b) with a narrow range of variability. This is also clearly shown in [39], due to the intrinsic anatomy characterization of TT and TC.

E2 showed the lowest performance for all brain planes and with all the considered percentages of training patients. This is probably due to the challenges of performing self-supervision from a dataset (i) with low inter-class variability and (ii) small in size (1527 images in total), which may confuse similarity loss during training. E1 showed good performance, reinforcing the concept that knowledge obtained from a bigger dataset could also be beneficial in a medical image contest [248, 249, 250].

E3 and E4 showed consistently high $F1$ for all brain planes and all the percentages of training images. Figure 4.5 shows the confusion matrices obtained with E4 and different percentage of training patients. Good performances are always reached when considering TC, having the plane unique characteristics (e.g. presence of cerebellum and cisterna magna) as opposed to TT and TV. As regard TT and TV, since these planes are very close to each other, it is more likely for them to be misclassified one with the other.

No significant differences were found when comparing E1, E2, E3, and E4 (ANOVA test p -value > 0.05) on the various percentages of training patients. However, this is not in accordance with the qualitative results obtained with CAM. It is in fact clear

from figure 4.6 how E4 experiment influences the way the architecture looks at the image to make a prediction.

Figure 4.6 shows sample visual results with CAM for Z-brain. It is worth noticing that, despite E4 having comparable numerical performance with respect to E3 and E1, CAM maps show that the dual-task training allows ResNet to focus on relevant anatomical structures in the image, following what clinicians do in clinical practice. TC plane permits in fact the evaluation of the cerebellum and cisterna magna, the two zones in which E4 focuses most. As regard TT and TV, they share almost the same structures except for the thalami for TT and the atrium for TV (see Fig. 4.1), the two zones in which E4 pays more attention (see Figure 4.6 b-c).

In Fig. 4.7, we show a visual confusion matrix with sample images along with CAM. TC is correctly classified when the network focuses on the cerebellum but is misclassified as TT when the network seems to spot a region resembling thalamus. Regarding papers in literature, the work that is more similar to ours is [65]. Starting from a dataset which consists of 13 + 1 (background) classes with different inter class variabilities (from brain to abdomen), they state that SSL is effective in separating distinct categories but can decrease performance in classes that are prone to overlap. Our training strategy differs from theirs: in order to evaluate this relationship, we decided to consider two separated datasets, one with high inter class variability (brain, thorax, maternal cervix and femur) and the other with low inter class variability (TC, TT and TV). This allows us to judge individually the behavior of the two datasets as opposed to [65]. While it is reasonable that similar classes are more prone to be mis-classified when trained with other different classes, a separated SSL training on these categories may benefit the classification.

4.5 Conclusion & Future Directions

In this work, we analyzed the impact of contrastive learning for standard plane detection in US fetal images. As clearly stated in Sec. 2, semi-supervised approaches are not completely exploited in this domain. The experimental results showed that end-to-end dual-task training including a pretext-task based on contrastive learning and a classifier may impact positively the performance in lower inter class variability datasets over standard transfer learning approaches. This finding is not trivial: self-supervised approaches are often used when a large number of unlabeled data is available, which is not representative of US datasets, though, paving the way for its possible usage even in this field. We are aware that contrastive learning exists in many different flavours, each with its own strengths and weaknesses. In our future experiments, we plan to explore different contrastive learning techniques as well as investigating the impact of contrastive learning for other tasks of fetal US image processing, including biometry estimation and computer-assisted diagnosis. Even though authors' aim is to evaluate the aforementioned methodologies when dealing with high and low intra class variability and, thus, assess the feasibility of contrastive

Chapter 4 On the use of contrastive learning for standard-plane classification

learning in this context, the dataset asset doesn't allow to mimic what is currently done in clinical practice: temporal dimension and background samples are in fact needed (refer to Sec 2.3.1). With the aim of translating these findings for fetus clinic examination, the aforementioned issues should be addressed, taking inspiration from other US works such as in [245, 63, 247], in which the temporal dimension is taken into account. Moreover, to better generalize the obtained results, a more extensive evaluation should be done, considering how these findings vary according to other different standard planes (e.g. all cardiac views, FFSP, etc).

Chapter 5

A deep-learning framework for metacarpal-head cartilage-thickness estimation

This Chapter describes a new method to automatically measure cartilage-thickness from metacarpal-head US images.

Cartilage damage is one of the most relevant determinants of physical disability in patients who suffer from Rheumatoid arthritis (RA), a chronic disease characterized by erosive symmetrical polyarthritis. Cartilage damage is nowadays assessed by clinicians, which manually measure cartilage thickness in ultrasound (US) imaging. This poses issues relevant to intra-and inter-observer variability. Relying on the acquisition of metacarpal-head US images from 38 subjects, this chapter addresses the problem of automatic cartilage-thickness measurement by designing a new deep-learning (DL) framework.

The framework consists of a Convolutional Neural Network (CNN), responsible for regressing cartilage-interface distance fields, followed by a post-processing step to delineate the two cartilage interfaces from the predicted distance fields and compute the cartilage thickness.

This work is the first to address the problem of automatic cartilage-thickness estimation in US rheumatological images with DL, paving the way for future research in the field. From a clinical perspective, the proposed framework proved to be a valuable tool to support the clinical routine increasing the reproducibility of cartilage thickness measurements.

A journal paper on this work has been published as [166]: Fiorentino, M. C., Cipolletta, E., Filippucci, E., Grassi, W., Frontoni, E., and Moccia, S (2021). A deep-learning framework for metacarpal-head cartilage-thickness estimation in ultrasound rheumatological images. *Computers in Biology and Medicine*, 141, 105117.

5.1 Introduction

Rheumatoid arthritis (RA) is a common systemic inflammatory autoimmune disease characterized by erosive symmetrical polyarthritis, which may severely impair physical

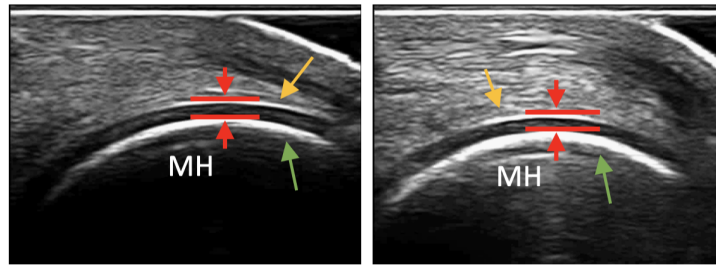


Figure 5.1: Two samples of thickness measurement of metacarpal head (MH) cartilage by US. Yellow arrow: chondrosynovial interface. Green arrow: osteochondral interface.

function and quality of life [251]. RA is frequently associated with joint destruction, involving both cartilage and bone damage, even though irreversible physical disability has been more clearly associated with cartilage damage than bone erosion [252, 253, 254]. Currently, conventional radiography is the most widely adopted imaging technique to assess joint damage in the clinical practice [255]. However, it has been shown that other techniques, such as US and magnetic resonance imaging, are more accurate to detect joint damage in RA [256]. Being radiation free and cost-effective, US is in fact one of the most promising techniques for the early detection and monitoring of both bone erosion and cartilage damage in RA [257, 258, 259].

In US images, the normal appearance of hyaline cartilage is characterized by a homogenous hypoanechoic layer delimited by two regular, sharp, and bright margins (the outer or chondrosynovial one, and the inner or osteochondral one). To date, two main methods have been proposed for the assessment of cartilage damage by US: semi-quantitative and quantitative methods. Semi-quantitative scoring systems are based on a qualitative and subjective evaluation of the hyaline cartilage [260, 261, 262]. In 2019, the Outcome of Measure in Rheumatology Ultrasound Taskforce published a three-semiquantitative score: grade 0 = normal cartilage, grade 1 = minimal change such as focal thinning or incomplete loss of cartilage, grade 2 = severe change such as vast thinning or complete loss of cartilage [263]. The major advantage of a semi-quantitative scoring system is its simplicity and speed. Quantitative-based scoring systems allow on the other hand to accurately measure the cartilage thickness and to monitor its changes [260, 261, 259]. During a quantitative-based scoring evaluation, cartilage thickness is measured by including the chondrosynovial interface but not the osteochondral one while keeping the line of measurement perpendicular to the subchondral bone surface. This process is shown in Fig. 5.1 for two metacarpal-head images [264]. This is a time consuming procedure (e.g., a comprehensive US evaluation from the II and V fingers of the hands takes at least 6-7 minutes), and may be prone to inter- and intra-observer variability, especially for observers with limited experience [260]. The availability of a software that allows automatic and real-time measurements during the US assessment could overcome this limitation.

In order to automatize the measurement process and improve inter- and intra-observer reliability, computer-assisted approaches have been developed in the years (Sec. 5.1.1). The most common challenges that have to be tackled in the observation process include presence of speckle noise, intensity in-homogeneity and high intra- and inter-subject variability due to individual variation of anatomical structures. These particular challenges have hampered the translation of the developed methodology in the actual clinical practice so far. During the last decade, artificial intelligence (AI) has become a tool in translational medicine, changing the way clinicians help patients [265]. In fact, the objective evaluation provided by AI algorithms may derive useful insights in order to better understand the mechanisms of a disease, by identifying features or details that may seem insignificant to the human eye [266]. For the aforementioned reasons, deep learning (DL) algorithms have been recently gaining relevance in aiding in US image analysis [1, 267, 3] since they are able to directly extract complicated information from input images. Convolutional neural networks (CNNs) have already shown promising results for cartilage segmentation [268, 269, 270]. This suggests that DL could potentially help tackling the challenges of automatic cartilage-thickness estimation. Nonetheless, to our knowledge, the potential of DL in the field hasn't been unlocked yet. This may be probably due to the lack of publicly-available annotated datasets in the field.

5.1.1 Related work

Considering cartilage thickness as an indicator for cartilage health, a number of approaches has been developed over the year with the intent of automatizing the thickness measurement and support clinicians in the clinical practice.

In [271], knee joint cartilage thickness is evaluated by enhancing cartilage contours by means of a multipurpose beta optimized recursive bi-histogram equalization method. The work is intended to overcome the limitation of traditional adaptive histogram method by integrating information coming from brightness shift and contrast enhancement. However, the approach strongly relies on the chosen setting which, in turn, cause the enhancement of soft-tissue interfaces, which could affect cartilage segmentation and thickness measurement.

As a prior for cartilage thickness estimation, the work in [272] performs knee cartilage segmentation using the locally statistical level set method which is based on shape and appearance terms failing to robustly model all variation present in the visual object of interest [273]. Also the framework proposed in [274] relies on cartilage segmentation. The segmentation, performed using a random-walker algorithm, takes place after an enhancement which is bone by means of local-phase image features and dynamic programming. Cartilage thickness is measured manually from the segmented region. The work in [274] is expanded upon in [275] in order to lower the need for manual intervention, by calculating the thickness as the Euclidean distance map from the segmented cartilage region. Despite the promising results, the segmentation

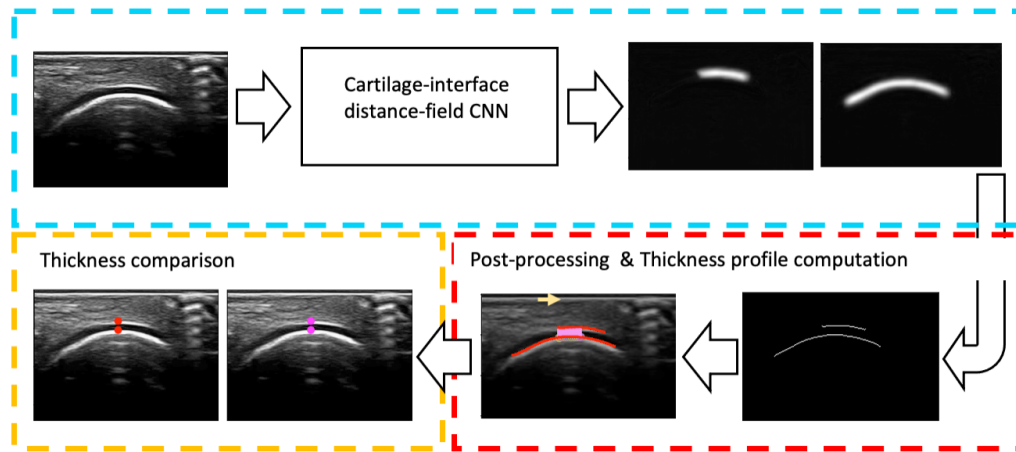


Figure 5.2: Proposed framework pipeline. A convolutional neural network is trained to predict the distance field from each of margin of the interfaces taken into account for thickness computation (lightblue box). A threshold operation is computed to extract the superficial margin of the chondrosynovial and the superficial margin of the osteochondral profiles from distance field. Cartilage thickness is then computed between corresponding points in the upper and lower interfaces, from the left-end (yellow arrow) to the right-end of the upper interface (red box). The obtained thickness measurement (purple points) is compared to that obtained by an expert clinician (red points).

of the cartilage strongly depends on hand-crafted features which vary as the anatomy of the subject varies, posing issues relevant to intra- and inter-subject variability.

DL may potentially offer a solution to tackle the challenges of cartilage-thickness estimation in US images since its algorithm training doesn't depend on domain expertise or hard core feature extraction, but the lack of annotated images has hampered the development of DL algorithms, so far. Following up on this consideration, in this work, we propose a framework based on CNNs for performing automatic cartilage-thickness measurement from US images acquired in the actual clinical practice.

5.1.2 Contribution

In this paper, we address the problem of automatic cartilage-thickness estimation in US rheumatological images, by designing and developing a new DL framework. The problem of cartilage-thickness estimation is tackled at first as a cartilage-interface delineation problem. Then, thickness is computed as the distance between points belonging to the superficial margin of chondrosynovial interface and the superficial margin of the osteochondral one.

To delineate the cartilage interfaces, one possible strategy could be to classify each image pixel as background or interface. However, the area covered by the interfaces is

small and narrow, and this would pose overfitting issues when training a classification network. Recently, researchers in other fields of medical-image analysis [1, 2, 276] have shown how edge delineation problems may be modelled as heatmap-regression tasks. In this work we follow this paradigm and regress the cartilage-interface distance fields via CNNs.

The contribution of this paper could be summarized as follow:

- Development of a new framework that is intended to estimate cartilage thickness from US images automatically. The framework relies on a CNN that regresses the superficial margin of the chondrosynovial interface and the superficial margin of the osteochondral interface via distance fields.
- Collection of the first dataset for DL applications in the field of metacarpophalangeal US image analysis for RA diagnosis. This dataset, acquired in the actual clinical practice, is collected from 38 subjects, with a mean of 24 (standard deviation ± 7.4) images per subject. The dataset is annotated by an expert clinician with the value of cartilage thickness.
- Comprehensive experimental validation, including the estimation of intra-observer variability.

Compared to what is currently done in the clinical practice, where the cartilage thickness is measured via only two points, the proposed framework produces a cartilage-thickness profile, i.e. the thickness is computed for each point-pair in the two interfaces. This provides the clinicians with additional quantitative data that can be used to perform early-stage detection of cartilage damage.

In this work, we focus our experiments on US metacarpal-head images, since metacarpal head is one of the most involved areas in RA patients [277]. Nonetheless, we believe that the proposed framework could be equally applied to other anatomical locations, such as femur cartilage.

5.2 Methods

The proposed framework for cartilage-thickness estimation from US images is shown in Fig. 5.2. A CNN is responsible for the regression of the cartilage interfaces via distance fields (Sec. 5.2.2). The CNN is followed by a post processing step to highlight the cartilage surfaces from the predicted distance fields. Cartilage thickness is computed by measuring the distance between each point in the upper interface and its corresponding point in the lower one (Sec. 5.2.2). This approach follows what is currently being done in clinical practice, where, however, only two points, one for the superficial margin of the chondrosynovial interface and one for the superficial margin of the osteochondral one, are taken into account by the clinicians.

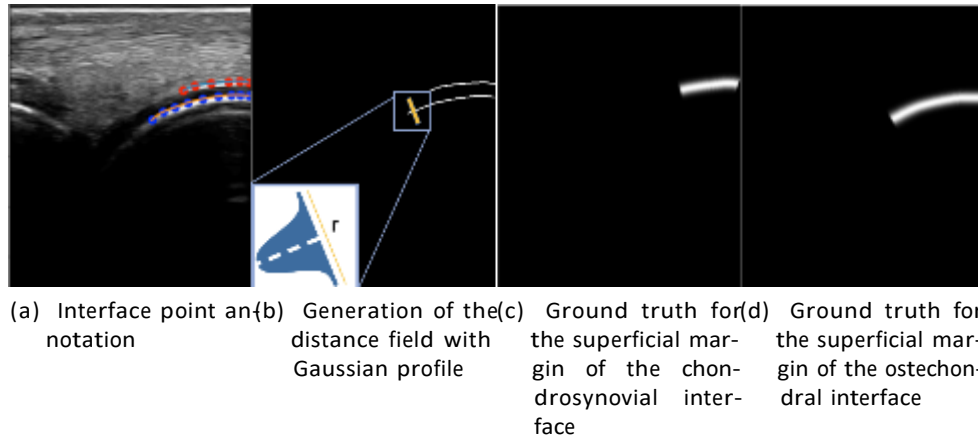


Figure 5.3: Ground-truth generation for distance fields. The raw image is annotated by drawing points along the interfaces of interest. (a) Cartilage interfaces delineation is obtained by cubic spline interpolation. The distance field is built by mounting a Gaussian distribution (length = r) in the line that is normal to the tangent line at each cartilage point (b). This is done for both the upper (c) and lower (d) interfaces, obtaining a double channel ground.

5.2.1 Data preparation

To develop our framework for cartilage thickness estimation, we rely on two ground-truth annotations:

1. Distance-field ground-truth images, which are used to train our regression CNN
2. Cartilage thickness annotations, which are made just like in the actual clinical practice (see Fig 5.2, Thickness comparison). Specifically, an expert clinician selects a point in the superficial margin of the chondrosynovial margin and draws the line perpendicular to the superficial margin of the osteochondral interface from that point.

To build the ground-truth images, we start from the manual delineation of the cartilage interfaces. The interfaces are annotated at first by manually selecting points along the boundaries of each interface (Fig. 5.3a). The annotation is performed under the clinician supervision using LabelMe [278]. Cubic spline interpolation is used to obtain the cartilage manual delineation from the annotated points. In order to build the distance-field ground truths from the manual delineation, we were inspired by [2]. For each point in each of the two interfaces, the normal line to the interface is found (Fig. 5.3b). A Gaussian profile (with length r and standard deviation $r/2$) is drawn along the normal line. A two-channel (one for the superficial margin of the chondrosynovial interface and one for the superficial margin of the osteochondral one) ground truth is built to avoid overlap among the interface distance fields and to separate the two surfaces without the need of any post-processing (Fig. 5.3c-d).

Table 5.1: Cartilages' regression network architecture with Conv2D = convolution, MaxPooling = max pooling, Upconv = up convolution, Concat = concatenation

	Operator	Kernel dimension	Dilation rate	Filters
Encoding path	Conv2D	3x3	1	64
	Conv2D	3x3	2	64
	MaxPooling	2x2	-	-
	Conv2D	3x3	1	128
	Conv2D	3x3	2	128
	MaxPooling	2x2	-	-
	Conv2D	3x3	1	256
	Conv2D	3x3	2	256
	MaxPooling	2x2	-	-
	Conv2D	3x3	1	512
	Conv2D	3x3	2	512
	MaxPooling	2x2	-	-
Bottleneck 2x	Conv2D	3x3	1	1024
Decoding path	Upconv	2x2	-	512
	Conv2D	2x2	1	512
	Concat	-	-	-
	Conv2D	3x3	1	512
	Conv2D	3x3	1	512
	Upconv	2x2	-	256
	Conv2D	2x2	1	256
	Concat	-	-	-
	Conv2D	3x3	1	256
	Conv2D	3x3	1	256
	Upconv	2x2	-	128
	Conv2D	2x2	1	128
	Concat	-	-	-
	Conv2D	3x3	1	128
	Conv2D	3x3	1	128
	Upconv	2x2	-	64
	Conv2D	2x2	1	64
	Concat	-	-	-
	Conv2D	3x3	1	64
	Conv2D	3x3	1	64
	Conv2D	3x3	1	2

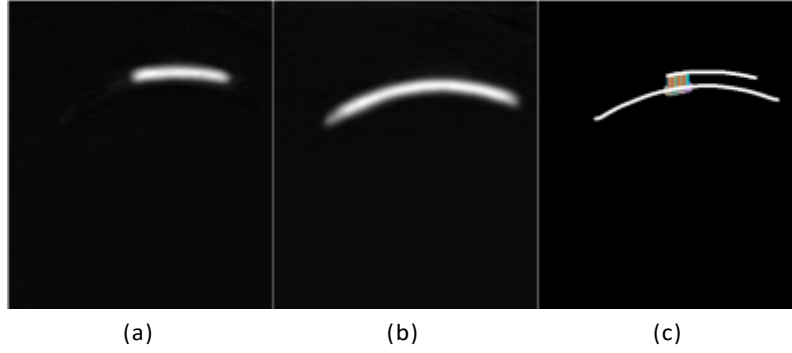


Figure 5.4: Post-processing procedure. The predicted distance fields for the upper (a) and lower (b) interfaces are thresholded and skeletonized (c). Cartilage thickness is calculated as the Euclidean distance between the corresponding points of the skeletonized interfaces (c).

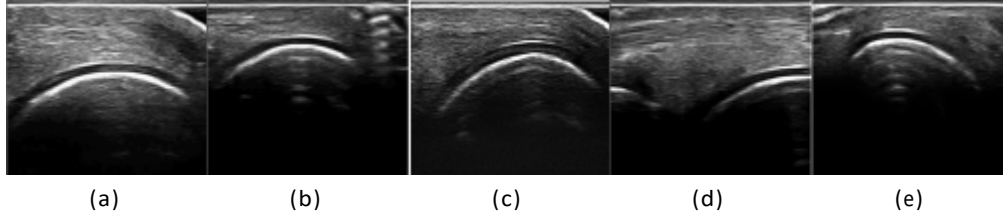


Figure 5.5: Challenges in the dataset include US artifacts (b-e), and anatomical (a-c) and cartilage position (d) variability.

5.2.2 Cartilage-surface distance-field CNN

To regress the cartilage-interface distance fields, we rely on a dilated-UNet network, as atrous convolutions have been demonstrated to be effective in boundary distance regression to segment kidney boundaries in US image [279]. Moreover, Unet is proven to be particularly effective with a limited dataset [229].

The CNN architecture is detailed in Table 5.1. It follows the Unet-structure of [229], which consists of a encoding, bottleneck and decoding paths.

The encoding path is made up of a sequence of one convolution (kernel size = 3×3 with dilation rate = 1, padding = 1), one convolution (kernel size = 3×3 with dilation rate = 2, padding = 1) and max pooling (stride = 2). This sequence is repeated 4 times. The number of filter is doubled at each iteration so as to learn more and more complex features. The bottleneck, which has the role to force the model to focus only on relevant information, connects the encoding path with the decoding path by means of two convolutional layers (kernel size = 3×3 , dilation rate = 1). The number of filters doubles from 512 to 1024. The decoding path consists of a sequence of one up convolution (kernel size = (2×2)) that halves the numbers of feature channels, followed by a convolution layer (filter = 2×2), which is linked via skip connections to the downsampling feature maps. Two convolutions in a row (kernel size = (3×3) ,

Table 5.2: Demographic characteristics of the subjects included in the study.

age (years, mean $\pm$ standard deviation)	54.6 $\pm$ 5.6
height (cm, mean $\pm$ standard deviation)	170.7 $\pm$ 9.5
handedness (right/left)	40/8
weight (kg, mean $\pm$ standard deviation)	72.4 $\pm$ 6.4
sex (female/male)	33/15
body mass index (kg/m^2 , mean $\pm$ standard deviation)	24.8 $\pm$ 1.7

dilation rate = 1, padding = 1) are successively used. This sequence is repeated 4 times. Every convolutional layer is activated by a rectified linear unit function and its output is batch normalized in order to stabilize the learning process. Finally, a convolution layer is used to reduce the number of filters to 2 in order to obtain the distance-field predictions.

Our CNN is trained using the stochastic gradient descent with momentum as optimizer, minimizing the mean squared error (MSE) loss function:

$$MSE(S, \hat{S}) = \frac{1}{N} \sum_{i=1}^N (S_i - \hat{S}_i)^2 \quad (5.1)$$

where S_i and $\hat{S}_i$ are the ground-truth and predicted distance field for the i -th training image, respectively, and N is the number of training samples.

Thickness profile computation

After obtaining the 2-channel distance-field prediction, Otsu's thresholding is used. The binarized images are skeletonized by means of Zhang's method [280] and re-sampled by means of a cubic spline interpolation to better smooth the interfaces. Then, a Gaussian filter is used to remove noise. To compute the cartilage thickness, for each point in the upper interface, the perpendicular to the surface is drawn until it intersects with a point in the lower interface. The cartilage thickness is then computed as the Euclidean distance between each point pairs (Fig. 5.4c).

5.3 Experimental setup

5.3.1 Dataset

To develop and evaluate the proposed framework, we collected a dataset from 38 subjects, approved by Comitato Etico Regione Marche id CERM 383/2019. In Table 5.2, information about subjects' age, height, handedness, weight, sex and body mass index is reported. Inclusion criteria were:

- No diagnosis of inflammatory/degenerative arthropathies

- No joint pain [visual analogue scale (VAS) $\leq 10/100$] and/or analgesic or non-steroidal anti-inflammatory drugs' intake in the four weeks prior to the visit
- Age > 18 -year-old
- No signs of osteoarthritis of metacarpophalangeal, proximal or distal interphalangeal joints at physical examination

Being the first attempt at exploring this kind of approach for cartilage thickness estimation, only healthy people were considered for this experiment, with the full intent to later extend this new suggested approach to subjects with RA.

US examinations were performed with a MyLab Class C (Esaote, Genoa, Italy), equipped with a very high frequency broadband linear transducer (10-22 MHz) by a sonographer with five years of experience. To increase the reproducibility, grey scale setting parameters were standardised (B-mode frequency: 22 MHz, master gain: 70%, mechanical index: 0.3, dynamic range: 12, depth 15 mm, focus position at the area of interest) and maintained unchanged during all the US examinations. [260]. The hands were placed on a table, with the metacarpophalangeal joints in maximal flexion ($>60^\circ$) [260, 263, 253]. The metacarpal heads from the 2nd to the 5th finger of both hands were scanned on the dorsal aspect from radial to ulnar and from proximal to distal sides to ensure the maximal exploration of the hyaline cartilage using the EULAR standard scans [281]. Particular attention was paid to ensure a perpendicular insonation of the cartilage surface [260, 263, 253].

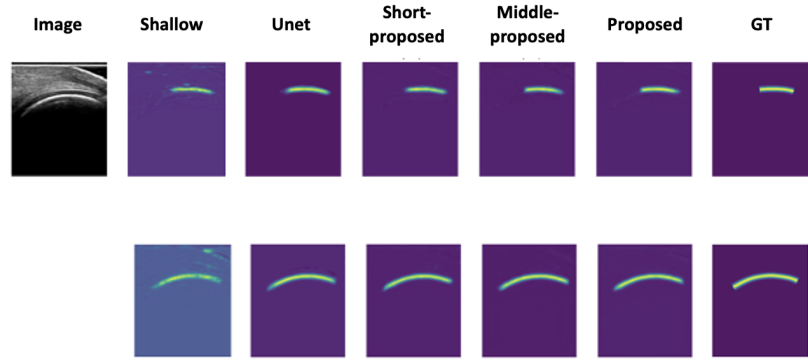
A mean of 24 (standard deviation = ± 7.4) images for subject was acquired, obtaining 904 images in total. Image size was 480x560 pixels, with a pixel size of 0.0310 [mm]. Challenges in the dataset (Fig. 5.5) included both typical challenges of US image analysis (reverberations, speckles, noise, unclear contours) and peculiar challenges of the specific anatomical district (i.e., different areas of metacarpal head cartilage, shape and appearance).

Training, validation and test sets consisted of 25 (number of images = 615), 5 (118) and 8 (173) patients, respectively. Dataset split for training and testing was performed at subject's level.

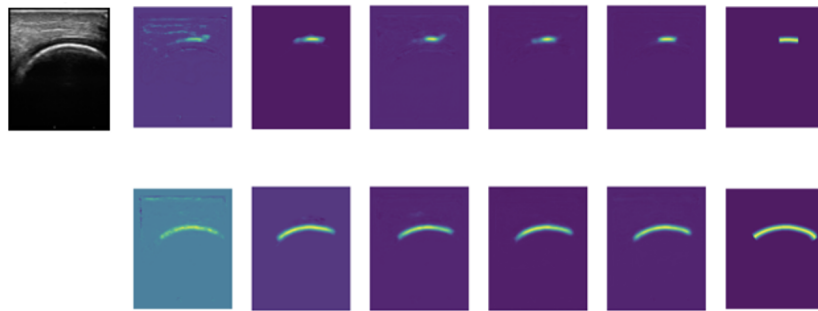
5.3.2 Parameter setting

Prior to feeding the CNN to obtain the distance-field regression, all training and validation images were resized to 240x280 pixels in order to limit both training time and memory requirements. Images were zero-padded from 240x280 to 240x288, to be compatible with the operations performed by the CNN. Mean intensity was removed from each image. The ground-truth images used for CNN's training were obtained with an r empirically set to 100 pixels and a Gaussian standard deviation of $r/2$. This allowed to fully cover the interfaces of interest. Once obtained the testing image

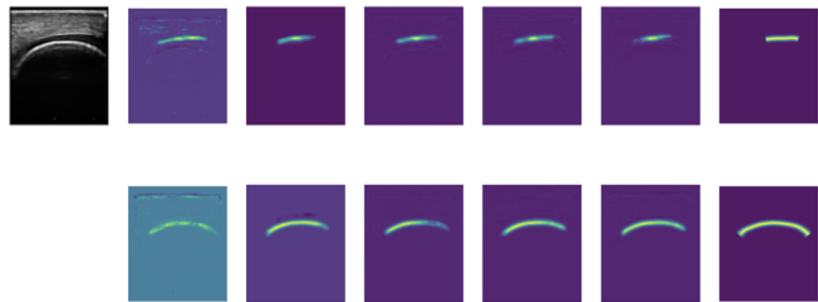
5.3 Experimental setup



(a) Sample 1



(b) Sample 2



(c) Sample 3



(d) Sample 4

Figure 5.6: Visual samples of results obtained from the ablation-study. Next to each image, shallow (1st column), Unet (2nd column), short-proposed (3rd column), middle-proposed (4th column) and proposed (5th column) prediction results along with ground true images (GT, 6th column) are shown. First row and second row related to each image⁹⁹ refer to upper and lower cartilage surface distance fields, respectively.

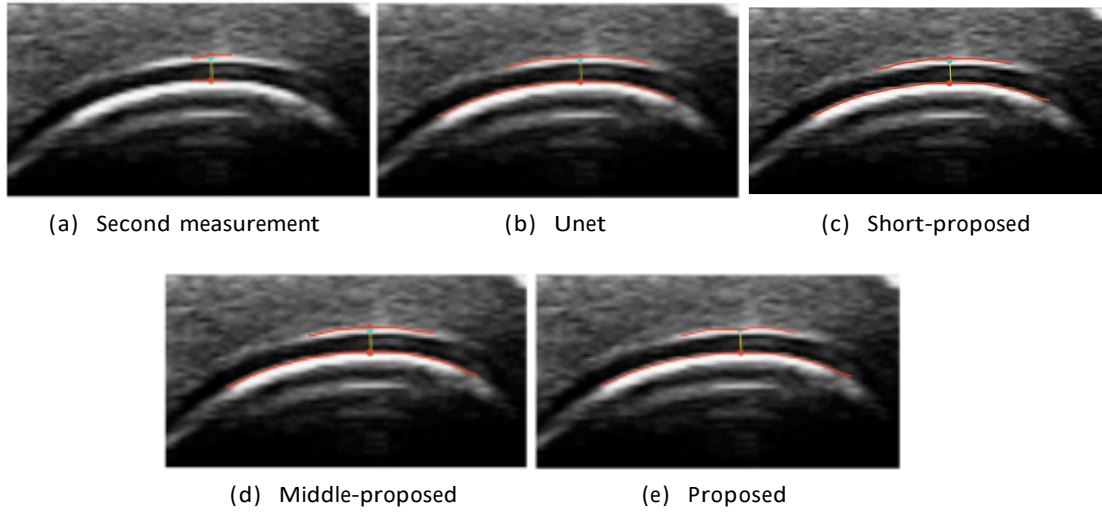


Figure 5.7: Visual examples of thickness measurement obtained by the architectures of our ablation study and by an expert clinician performing the measurement for the second time. The sea water points are the ones annotated by the clinician and used as ground true. Red points are the closer points to the sea water point, as estimated by each architecture of the ablation study and second clinician annotation. Images are zoom in to better appreciate results.

predictions, the latter images were scaled back to 480x560 pixels to evaluate the proposed approach in the real image dimension.

The CNN was trained minimizing the mini-batch stochastic gradient descent with momentum = 0.98. A learning rate = 0.001 for 1500 epochs was used. Batch size was set to 8, as a trade off between memory requirements and training convergence. On-the-fly data augmentation was performed during training by horizontal flipping and rotating (range $\pm 45^\circ$) images. The transformations were randomly applied at each training iteration. The best model was selected according to the lowest mean absolute error (MAE) obtained in the validation set:

$$MAE(M, \hat{M}) = \frac{1}{N} \sum_{i=1}^N |M_i - \hat{M}_i| \quad (5.2)$$

where M_i and $\hat{M}_i$ are the regression mask and predicted outcome of i-th training image, respectively, and N is the number of training samples. All the analyses were performed using Keras¹ on a NVIDIA RTX 2080TI, with a Xeon e5 CPU and 128 GB RAM.

¹<https://keras.io>

Table 5.3: Ablation study summary. Encoder-decoder structure, atrous convolutions and depth (number of layers) are the model characteristics taken into account.

	Encoder-decoder	Atrous convolutions	Depth
Shallow		X	1
Unet	X		4
Short-proposed	X	X	2
Middle-proposed	X	X	3
Proposed	X	X	4

5.3.3 Performance metrics

To assess the performance of our CNN network, we computed the *MAE* between the predicted distance-field and the distance-field ground truth in the test images.

In order to mimic what is currently done in clinical practice, obtaining a more comparable understanding of the potential of our findings, an expert clinician measured cartilage thickness in each of the testing image, as described in Sec 5.2.1. The following metrics were computed:

- Difference (*DF*) [mm]:

$$DF = CT_A \square CT_B \quad (5.3)$$

- Absolute difference (*ADF*) [mm]:

$$ADF = |CT_A \square CT_B| \quad (5.4)$$

with CT_A and CT_B are the cartilage thickness obtained by the proposed framework and computed by the clinician, respectively. *ADF* is used to determine accuracy of predictions, on the other hand *DF* is used to assess the over/underestimation of the thickness.

5.3.4 Ablation study & intra-clinician variability assessment

We carried out an ablation study to identify the best configuration of our framework. The effectiveness of the encoder-decoder structure of our CNN was investigated by training a shallow CNN (*Shallow*). The shallow CNN consisted of a sequence of one convolution (kernel size = 3x3 with dilation rate = 1, padding = 1) and one convolution (kernel size = 3x3 with dilation rate = 2, padding = 1), each followed by a batch normalization layer. The sequence was repeated 4 times (with 64, 128, 128 and 64 filters for the first, second, third and forth repetition, respectively). The effect of the CNN depth on the distance-field regression performance was investigated, too. We trained two additional architectures, which were obtained by removing the deeper

Table 5.4: Ablation study comparison in terms of mean absolute error between distance field architectures predictions and distance field ground truth images.

	Mean Absolute error
Shallow	0.014 ($\pm$ 0.003)
Unet	0.0041 ($\pm$ 0.001)
Short-proposed	0.0054 ($\pm$ 0.001)
Middle-proposed	0.0051 ($\pm$ 0.001)
Proposed	0.0050 ($\pm$ 0.001)

layers of the proposed architectures, stopping at 512 filters for *Middle-proposed* architecture and 256 filters for *Short-proposed* architecture. Dilation rate was also evaluated by training a standard *Unet* architecture without atrous convolutions. Table 5.3 shows a summary of our ablation study. For fair comparison, we always used the same dataset split and training settings, as well as the same computational hardware.

To assess the intra-observer variability, the expert sonographer who provided the ground truth thickness measurement repeated the measurement a second time.

5.4 Results

The performance comparison in terms of *MAE* is reported in Table 5.4. The highest $MAE = 0.014 (\pm 0.003)$ is obtained by Shallow architecture, followed by the Short-proposed one. Middle-proposed and proposed architectures performed comparably well with a $MAE = 0.0051 (\pm 0.001)$ and $0.0050 (\pm 0.003)$, respectively. The best $MAE = 0.0041 (\pm 0.003)$ is instead reached by Unet. In Fig. 5.6, samples of cartilage regression distance field predictions are shown. Results are obtained by shallow (first column), Unet (second column), short-proposed (third column), middle-proposed (fifth column) and proposed (fifth column) architectures, respectively. The distance field ground truth is shown too. The first and second row correspond to upper and lower cartilage surface, respectively.

The performance comparison in terms of *DF* and *ADF* is reported in Table 5.5. We decide to not show Shallow architecture performances because of the poor results obtained in *MAE* calculation. The best performance in term of *ADF* is obtained by the proposed architecture, followed by its shorter versions (short-proposed and middle-proposed). An $ADF=0.039 (\pm 0.029)$ mm is instead obtained by Unet, which was the best in *MAE* calculation. As regard *DF*, the proposed architectures obtained the lowest values as opposed to unet and clinician. Figure 5.7 shows visual examples of thickness measurement performed by ablation study architectures, which

Table 5.5: Ablation study & intra-clinician variability assessment in terms of absolute difference and difference. We report the mean for each metric, with standard deviation in brackets. Metrics are calculated between the cartilage thickness obtained by the architecture (or for the second time by the same clinician) and clinician manual annotation.

	Difference [mm]	Absolute Difference [mm]
Clinician (2nd time)	-0.009 (0.045)	0.036 (0.028)
Unet	0.010 (0.048)	0.039 (0.029)
Short-proposed	-0.002 (0.045)	0.035 (0.028)
Middle-proposed	0.003 (0.041)	0.035 (0.026)
Proposed	-0.004 (0.041)	0.032 (0.026)

are compared to the one performed by the sonographer. Intra-variability of the measurement is also shown. The sea water color points are the ones annotated by the clinician used as ground true optimal location for thickness calculation. The red points are the closest to the sea-water colored ones which were first annotated by the clinician. The same red points were estimated by ablation study architectures and second clinician annotation.

5.5 Discussions

Preventing or halting cartilage damage is of paramount importance in RA. Cartilage thickness measurement using US is a promising method to early detect and monitor cartilage damage. In daily clinical practice, such measurement is performed manually by sonographer, posing issues related to intra- and inter-observer variability. Different computer-assisted approaches have been developed over time with the aim of attenuating such issues, but none of them has fully exploited the potential of DL, yet.

To unlock the potential of DL in the field, we collected a dataset of 38 subjects (with a mean of 24 images each) that underwent a US assessment of the metacarpal head cartilage, facing challenges related to US artifacts (such as reverberations and speckles) and anatomical variability of the investigated anatomical structures. In light of the considerable effort from the clinician side in providing, not just images acquisition, but more importantly the corresponding image-level annotations, we had to find a trade off between such effort and dataset size. With this considerations in mind, data were acquired taking into account as much variability as possible (Fig. 5.5). Relying on this dataset, we developed a DL framework for automatic cartilage thickness estimation. The framework consisted of a CNN that regresses distance field from the cartilage interfaces. In fact, directly segmenting lines is a well known problem since the number of pixels belonging to each interfaces is small with respect

to the total image dimension, which makes challenging minimizing the most common segmentation losses (e.g. DSC, categorical cross entropy). The cartilage thickness was computed operating a post-processing on the predicted distance fields. As a result, a cartilage thickness profile is provided to the clinician. The proposed framework achieved encouraging results with a mean $ADF = 0.032$ (standard deviation = ± 0.026) mm, $DF = -0.004$ (± 0.041) mm, proving to be an accurate method for cartilage thickness measurement.

From our ablation study, Unet, short-proposed, middle-proposed and the proposed architecture performed quite well in estimating the distance field of the upper cartilage surface in US images with contrasted cartilage surfaces (Fig 5.6, upper-left image), with a long, wide and curve upper interface cartilage. The shallow architecture performed slightly less precisely, although the error in the distance-field estimation did not affect the cartilage-thickness measurement.

As the upper left interface was less visible (Fig. 5.6a), shallow architecture predictions deteriorated. Short-proposed architecture predictions deteriorate too, but not drastically, as opposed to Unet, middle-proposed and proposed that still performed comparably well. This may be attributed to the power of encoder decoder structure which has the ability to extract use-full features, whose expressibility increases exponentially with the depth [282]. For this reason, the shallow architecture, which consists of a sequence of convolution, fails to identify surface boundaries with precision, and the depth of the short-proposed is not as such to precisely locate distance field pixels on image.

Images in which the upper cartilage surface appeared flat, thus not characterized by the typical curvature, were particularly difficult for any architecture (Fig. 5.6c). This is due to the fact that this kind of images are few compared to the whole dataset, typically characterized by curved upper surfaces.

Regarding lower interface predictions, Unet, short-proposed, middle-proposed and proposed architectures performed quite well with some exception by short-proposed (Fig. 5.6c-2d). Shallow architecture had difficulties even in the lower interface regression.

The best MAE was reached by the standard Unet (Table 5.4). However, even though it is able to regress the distance field in the side portions of the cartilage interfaces as opposed to the proposed architecture, upper surface predictions appear less precise (see Figure 5.7b-c). Atrous convolutions offers, in fact, a way to control the field-of-view, finding the best trade-off between accurate localization (small field-of-view) and context assimilation (large field-of-view), integrating therefore local and general information together. This permits the architecture to better focus on edges. As a matter of fact, Unet architecture covers a larger portion of the surfaces but seems to be less precise in delineating it, obtaining an $ADF = 0.039$ (± 0.028) mm (Tab. 5.5). Short-proposed and Middle-proposed perform comparably well with a $ADF = 0.035$ (± 0.028) mm and 0.035 (± 0.026) mm, respectively. As the depth increases, the proposed network becomes more and more

able to discriminate cartilage boundaries. This is particularly visible by looking at Fig. 5.7: points annotated by clinician and points estimated by proposed architecture overlap completely, proving to be a valuable tool to cartilage thickness calculation. The lowest DFs are obtained by proposed architectures proving to be less prone to over/underestimate the measurement.

In fact, whatever the choice of the network is, the pipeline of the entire work seems to be particularly promising, except for those architectures that do not have a encoder-decoder structure. Evaluating other recent architectures (attentions, multi-scale feature concatenation, siamese learning) is out of the scope of this work, which is indeed focused on presenting the entire pipeline for estimating cartilage thickness in US images.

The ADF of $0.036 (\pm 0.028)$ mm obtained during the intra-clinician variability analysis was higher than the ADF obtained with the proposed framework ($0.032 (\pm 0.026)$ mm). This suggests that our framework may be a true support in the clinical routine with a view to make RA assessment more reproducible. As opposed to state-of-the-art approaches, our framework doesn't rely on domain expertise or hard core feature extraction, making our framework very flexible to be adapted to new data.

Limitations in our experimental protocol include 1) poor dataset size 2) acquisition from the same device and annotation performed by the same clinician 3) inclusion of healthy subjects. With a view to translate the proposed pipeline in the actual clinical practice, these limitations must be overcome. However, we believe that this dataset, consisting of 38 subjects, whose size is similar to other datasets in literature², is a valuable initial step to stimulate further research in the field. In fact, US image challenges are getting of increasing interest among researches, enough to be considered almost every year in the Gran Challenge of MICCAI (e.g. A-AFMA³ and HC18 challenge [41]).

A way to directly delineate the upper and lower surfaces could also be exploited in order to not rely on thresholding methods. This could be done by adding a CNN to train cartilage surfaces end-to-end with distance-field CNN [279]. In addition, the proposed work contributes to increase the research in the field of cartilage damage and places itself among the latest approaches to the matter [283, 284, 285].

5.6 Conclusions

In this work, we suggest an innovative framework for cartilage-thickness measurement in US metacarpophalangeal images by acquiring the first metacarpophalangeal dataset suited for DL analysis. Our framework relies on distance fields to accurately delineate cartilage interfaces. With our framework, we achieved an $ADF = 0.032 (\pm 0.026)$ mm from the annotation performed by an expert clinician, while our

²<https://fetreg2021.grand-challenge.org/>

³<https://a-afma.grand-challenge.org/>

experiment on intra clinician variability led to an $ADF = 0.036 (\pm 0.028)$ mm. This proved our framework to be a valuable tool to support clinicians. This work is among the first attempts at exploiting DL in cartilage thickness measurement and has great potential to be translated in the clinical practice.

Chapter 6

Conclusions

The general goal of this PhD thesis is to develop automatic and reliable algorithms based on DL with the aim of improving the understanding of US images. Although US image analysis is growing rapidly, the skill and experience of the clinician is still at the core of US image analysis. The clinician's decision-making process is still subjected to personal biases and it takes a considerable amount of time to assess various factors such as biometric parameter estimation or identify specific plane search. Therefore the methods in this paper were developed through systematic analysis and modeling of US data with the specific aim of i) supporting the clinician's decision-making process through trained algorithms and ii) providing a reliable method in areas where standard procedures are lacking.

In Chapter 2, an extensive review of fetal US DL algorithm applications is conducted in order to increase the understanding of what is currently missing in the field, underlying limitations and open issues of each surveyed task (e.g. SPD, ASA, BPM and other tasks). Chapter 3 tackled the problem of HC parameter estimation, a well known research field of the domain, by exploiting a boundary distance field approaches. The experimental results in the only publicly available dataset of the field have demonstrated that the proposed methods could automatically extrapolate HC measurement with promising performance, overcoming DL pixel classification networks. Chapter 4 exploits SSL approaches for SPD with constrastive learning paradigm, a yet unexplored domain in this field. Collecting and sharing annotated datasets for DL algorithm development is a well-known problem for medical-image analysis community, and adapting SSL in the medical image analysis contest is not trivial.

With Chapter 5, cartilage thickness estimation is performed in order to evaluate the progression of RA, confirming that DL algorithms may potentially have clinical implication apart from to streamline the clinician's workload.

6.1 Thesis Contributions

The PhD research contributed to:

- *Introduce and discuss the most innovative and effective DL methods found in the literature for fetal US image analysis. An updated review may be a valuable and*

compact source of information for young researchers and a reference document for those already working in the field

This review presented a detailed analysis of a wide spectrum of the most investigated task, from SPD up to a number of emerging tasks. The availability of different datasets has increased the research in HC estimation field (thanks to the release of the HC18 challenge) along with SPD classification (i.e. zenodo dataset). The major performance improvement that DL algorithms may achieve, is in fact strongly connected to the availability of datasets, as a fair algorithm comparison could be performed. Despite the efforts of some International Organization and some researchers, the lack of available datasets in each sector is still something that has to be addressed. The acquisition of relevant annotations/labeling for these images is another open challenge: the lack of image annotation performed by multiple clinicians hampers the development of robust DL algorithms and the subsequent integration on device for commercial use. DL algorithms show good performance in a controlled evaluation setup, but this does not reflect the actual clinical practice, in which different US devices and probes are used. Other data acquisition challenges include: i) providing information about GA distribution, ii) inclusion of congenital anomalies for CAD systems training. In this perspective, researchers should work on creating balanced datasets, possibly exploring generative approaches or multicentric studies.

As for now, DL techniques have demonstrated to handle different challenges such as inhomogeneities, artifacts, poor contrast and intra- and inter-clinician acquisition and measurement variability, supporting clinicians in a wide range of clinical applications. When dealing with fetal analysis, algorithms have also to tackle challenges related to high variability (size and shape) of fetus among different GAs along with presence of shared anatomical structures among fetal standard planes (e.g. brain and heart planes). DL algorithms are also an essential tool in supporting segmentation process which is the primary step to estimate different parameters (e.g. area or volume) and avoid possible malformations. The choice of the DL architecture to use along with expert knowledge are two important keys for a good performance: innovation often happens through architectural changes.

Open issues of the field to address include i) the creation of a end-to-end approach able to correctly classify fetal planes and successively evaluate biometrics or fetus defect ii) semi, weak and self-supervised learning inclusion, even if there is a recent attempt to exploit such techniques in the field iii) adherence to ethics guidelines for trustworthy AI

- *Develop a new method which exploits a distance field strategy to HC measurement estimation*

US imaging has been widely used to assess fetal size and GA as well as detecting fetal growth abnormalities. Fetal biometry assessment represents the most common medical investigation undertaken in this regard and among different biometries, HC is measured. In the clinical practice, its measurement is performed manually, which increases inter-operator variability and inter-operator variability. Automatic and reliable HC delineation in US imaging data would improve precision and reliability of the measurement. A novel distance field network approach is proposed to delineate HC boundaries instead of using a pixelwise classification network which distinguishes fetal head pixels from background. Methods proposed (2step RCNN and Mask-R²CNN) are trained and evaluated using the HC18 challenge dataset, which allow a direct comparison with other DL approaches. Experimental results have demonstrated that distance field produces precise delineations of HC boundaries invariably from shapes, heterogeneous appearances and gestational age of fetus head. The end-to-end approach (Mask-R²CNN) has moreover the greater advantage of being more easily integrated into the device.

- *Evaluate the feasibility of self contrastive learning to train a semi-supervised fetal standard plane classifier*

For many years, DL algorithms have been largely dependent on good quality labeled data, a procedure that represents a major bottleneck in the overall training process, especially in US field in which data labeling is something that has to be done by expert personnel. One of the top priorities in the research community was the development of self-learning mechanisms that may take advantage of unstructured data. To solve the aforementioned problem, a lot of research has been conducted on self-supervised learning techniques capable of capturing subtle nuances in data. However, directly adapting these techniques in US domain is not trivial. First of all, when considering limited data annotation, to which extent using contrastive learning improves performance over standard transfer learning from large datasets of natural images? Second of all, should inter-class variability be taken into account to decide whether to use contrastive learning? Guided by these research questions, we studied the use of SSL applying self-supervision in a well-known problem like fetal standard plane detection, taking advantage of the only publicly available dataset of the field. Using SimCLR, a well-known framework, experiments showed that contrastive learning training is particularly advantageous in case of small inter-class variability, whereas it has comparable benefit to pretraining on natural images when working with a high inter-class variability dataset.

- *Develop a new framework for metacarpal-head cartilage-thickness estimation in US rheumatological images*

RA arthritis is the most common reumatic disease, resulting in joint swelling,

joint stiffness and musculoskeletal pain [286]. The longterm function disability of patient with RA is mainly due to joint destruction, involving both cartilage and bone damage. In clinical practice, cartilage and bone damage are assessed through conventional radiography even if US has proved to be the most accurate in the detection of joint damage in RA [256]. Other advantages include its cost-effectiveness and safety, making it one of the most promising techniques for the early detection of cartilage damage. As for now, DL in musculoskeletal US has received significant attention and it is expected to be pivotal in the development of next-generation cutting-edge US imaging systems, as stated in [9]. Different automatic or semi-automatic algorithms have been proposed over the year to estimate cartilage thickness but no one exploited DL approaches. This may be due to the objective difficulty to collect enough labelled data. The problem of automatic cartilage-thickness estimation is addressed in Chapter 5 by tackling it as a cartilage interface delination problem. A dataset consisting of the acquisition of metacarpalhead US images from 38 subject is used. The framework achieved encouraging results with a ADF of 0.031 ($\pm$ 0.026) against manual thickness annotation that was comparable to the intra-observer variability, proving the potential of DL in the field. This work is the first to address the problem of automatic cartilage-thickness estimation in US rheumatological images with DL, paving the way for future research in the field. From a clinical perspective, the proposed framework proved to be a valuable tool to support the clinical routine increasing the reproducibility of cartilage thickness measurements along with a new way to conduct cartilage measurement, considering its calculation between each point of the two interfaces.

6.2 Impact

DL technologies have the potential to outperform humans in some visual and auditory perception tasks, suggesting its application in medicine and healthcare, especially in US imaging. US is a clinician dependent device with unique challenges that include operator dependence, noise, artifacts, limited field of view and variability across different US manufacturers [287]. The results of this PhD work may impact positively the work of US sonographers during clinical procedures. While the use of US is becoming pervasive, advanced US image analysis techniques are still not incorporated on-device as opposed to CT or MRI. Reliance on the operator's ability is especially restricting: numerous medical service suppliers who are not imaging experts do not utilise US at the point of care because of an absence of expertise in attaining and deciphering pictures. Those who are familiar with US imaging still struggle with inter and intra-operator inconsistent decision-making. This inconsistency hinders the process of encouraging US clinical usage such as for serial oncological imaging, as US-derived tumour measurements end up not being accepted in many cancer drug trials. Automated US image analysis vows to assume a pivotal part in tackling these

issues. With the betterment of the patients' quality of life in mind, the procedures proposed in this PhD thesis will help in addressing some of the open problems by ensuring a more dependable decision-making process. The results achieved by the proposed methodologies pave the way to making US reliable enough to be used in clinical practice, though there's still much to be done (for instance multi-centric validation in clinical practice, a serious procedure which requires multiple partners with different skillsets).

6.2.1 Future Perspective

From the dissertation of this thesis, it is evident that DL has entered various application areas of medical US analysis. However, although DL techniques continuously update SOTA performance results for various application domains in medical US analysis, much remains to be done. As clearly pointed out by the vast majority surveys of the domain [14, 287, 1, 7], the higher data variability of the acquired dataset, the higher the performance improvement that can be achieved with DL. However, differently from other research areas in which datasets are large and publicly available (e.g., more than 1 million annotated multi-label natural images in ImageNet [288]), in the field of US, there is an objective difficulty in acquiring structured data along with clinicians annotations. Limited training data is a bottleneck for i) developing further clinical DL applications ii) integrating DL algorithms in clinical practice. However, it is hard to expect that US datasets will reach the appropriate size in a short-term future, pointing out the needing of finding different DL training strategies, which can be suited for smaller databases. [287] Transfer learning has been widely exploited and its effectiveness is also clearly shown in Chapter 4 and several other works [249]. Semi and weakly supervised learning have been successfully used [289] but their application is more suitable when larger sets of unlabeled data are widely available. In addition to the aforementioned techniques, other approaches which address small, annotated databases are the unsupervised learning ones, which exploit the reconstruction of data while learning a good representation, or learning similarities among data by means of contrastive learning loss. One of the first attempts at exploring contrastive learning loss for SPD is shown in Chapter 4, but not generalizing to other applications creates biases: even if grand-challenges are a great resource for benchmarking algorithms, these often address only a single application with the risk of overfitting to the specific dataset. Recent progress of self-supervised visual representations has, however, paved the way for applying these methodologies in domain adaptation applications [290, 291, 292], which may be exploited to integrate together the different publicly available datasets of the field and possibly develop end-to-end approaches that first identify the standard frame and then calculate the biometrics related to that image. Federated learning is a novel paradigm for data-private multi-institutional collaborations that could also be exploited to incentive collaboration among clinical center. It consists in the training of algorithms across

Chapter 6 Conclusions

multiple decentralized edge devices without exchanging private datasets. Another challenge that hampers the integration of DL algorithms for US image analysis is the need for algorithm interpretability. In fact, due to the DL black-box nature, it is often difficult to understand prediction results or being aware of possible biases [293] and an extensive research has to be conducted even in this regard. As a final remark, with the advance in technology and the rapid development of DL algorithms, the application of ethical principles and guidelines has become crucial. As also stated in [14], no attempts have been made in respecting and promoting ethical principles at least in the fetal US domain.

Acronyms

AC Abdominal Circumference

Acc Accuracy

AI Artificial Intelligence

AUC Area Under the ROC Curve

BPD Biparietal Diameter

CAD Computer-Assisted Diagnosis

CNN Convolutional Neural Networks

CRL Crown-Rump Length

CSP Cavum Septum Pellucidum

CTR Cardio-Thoracic Ratio

DL Deep Learning

DRL Deep Reinforcement Learning

DSC Dice Similarity Coefficient

ED Euclidean Distance

FASP Fetal Abdomen Standard Plane

FBSP Fetal Brain Standard Plane

FCN Fully Convolutional Networks

FFASP Fetal Face Axial Standard Plane

FFSP Fetal Facial Standard Plane

FFESP Fetal Femur Standard Plane

FL Femur Diaphysis Length

FLVSP Fetal Lumbosacral Spine Standard Plane

FV Fetal Ventriculomegaly

F1 F1-score

GA Gestational Age

GAN Generative Adversarial Network

Chapter 6 Conclusions

GRU-RCN Gated-Recurrent-Unit Recurrent Convolutional Network

HC Head Circumference

HD Hausdorff Distance

IoU Intersection over Union

ISBI International Symposium on Biomedical Images

ISUOG International Society of Ultrasound in Obstetrics & and Gynecology

KLD Kullback-Leibler Divergence

LVOT Left Ventricular Outflow Tract

LVR Lateral Ventricle Ratio

LSTM Long Short-Term Memory

MAE Mean Absolute Error

MICCAI Medical Image Computing and Computer Assisted Intervention

MVP Maximum Vertical Pocket

NAS Neural Architecture Search

NSS Normalized Scanpath Saliency

OFD Occipito-Frontal Diameter

Prec Precision

RA Rheumatoid arthritis

Rec Recall

RMSE Root Mean Squared Error

RNN Recurrent Neural Network

ROC Receiver Operating Characteristic

RVOT Right Ventricular Outflow Tract

Spec Specificity

SSD Single-Shot Detector

SSL Self-Supervised Learning

TC Trans-cerebellar

TT Trans-thalamic

TV Trans-ventricular

US Ultrasound

Bibliography

- [1] Shengfeng Liu, Yi Wang, Xin Yang, Baiying Lei, Li Liu, Shawn Xiang Li, Dong Ni, and Tianfu Wang. Deep learning in medical ultrasound analysis: a review. *Engineering*, 5(2):261–275, 2019.
- [2] Maria Chiara Fiorentino, Sara Moccia, Morris Capparuccini, Sara Giamberini, and Emanuele Frontoni. A regression framework to head-circumference delineation from US fetal images. *Computer Methods and Programs in Biomedicine*, page 105771, 2020.
- [3] Ruud J G Van Sloun, Regev Cohen, and Yonina C Eldar. Deep learning in ultrasound imaging. *Proceedings of the IEEE*, 108(1):11–29, 2019.
- [4] Aladin Carovac, Fahrudin Smajlovic, and Dzelaludin Junuzovic. Application of ultrasound in medicine. *Acta Informatica Medica*, 19(3):168, 2011.
- [5] Athanasios Voulodimos, Nikolaos Doulamis, Anastasios Doulamis, and Eftychios Protopapadakis. Deep learning for computer vision: A brief review. *Computational intelligence and neuroscience*, 2018, 2018.
- [6] S Kevin Zhou, Hayit Greenspan, Christos Davatzikos, James S Duncan, Bram Van Ginneken, Anant Madabhushi, Jerry L Prince, Daniel Rueckert, and Ronald M Summers. A review of deep learning in medical imaging: Imaging traits, technology trends, case studies with progress highlights, and future promises. *Proceedings of the IEEE*, 109(5):820–838, 2021.
- [7] Geert Litjens, Thijs Kooi, Babak Ehteshami Bejnordi, Arnaud Arindra Adiyoso Setio, Francesco Ciompi, Mohsen Ghafoorian, Jeroen Awm Van Der Laak, Bram Van Ginneken, and Clara I Sánchez. A survey on deep learning in medical image analysis. *Medical Image Analysis*, 42:60–88, 2017.
- [8] Jan Egger, Christina Gsaxner, Antonio Pepe, Kelsey L Pomykala, Frederic Jonske, Manuel Kurz, Jianning Li, and Jens Kleesiek. Medical deep learning—a systematic meta-review. *Computer Methods and Programs in Biomedicine*, page 106874, 2022.
- [9] YiRang Shin, Jaemoon Yang, Young Han Lee, and Sungjun Kim. Artificial intelligence in musculoskeletal ultrasound imaging. *Ultrasonography*, 40(1):30, 2021.

Bibliography

- [10] Gianluca Smerilli, Edoardo Cipolletta, Gianmarco Sartini, Erica Moscioni, Mariachiara Di Cosmo, Maria Chiara Fiorentino, Sara Moccia, Emanuele Frontoni, Walter Grassi, and Emilio Filippucci. Development of a convolutional neural network for the identification and the measurement of the median nerve on ultrasound images acquired at carpal tunnel level. *Arthritis Research & Therapy*, 24(1):1–8, 2022.
- [11] Mariachiara Di Cosmo, Maria Chiara Fiorentino, Francesca Pia Villani, Gianmarco Sartini, Gianluca Smerilli, Emilio Filippucci, Emanuele Frontoni, and Sara Moccia. Learning-based median nerve segmentation from ultrasound images for carpal tunnel syndrome evaluation. In *2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, pages 3025–3028. IEEE, 2021.
- [12] Philippe Burlina, Seth Billings, Neil Joshi, and Jemima Albayda. Automated diagnosis of myositis from muscle ultrasound: Exploring the use of machine learning and deep learning methods. *PloS one*, 12(8):e0184059, 2017.
- [13] Vincenzo Venerito, Orazio Angelini, Gerardo Cazzato, Giuseppe Lopalco, Eugenio Maiorano, Antonietta Cimmino, and Florenzo Iannone. A convolutional neural network with transfer learning for automatic discrimination between low and high-grade synovitis: a pilot study. *Internal and Emergency Medicine*, 16(6):1457–1465, 2021.
- [14] Maria Chiara Fiorentino, Francesca Pia Villani, Mariachiara Di Cosmo, Emanuele Frontoni, and Sara Moccia. A review on deep-learning algorithms for fetal ultrasound-image analysis. *Medical Image Analysis*, 83:102629, 2023.
- [15] Paolo Zafino, Sara Moccia, Elena De Momi, and Maria Francesca Spadea. A review on advances in intra-operative imaging for surgery and therapy: imagining the operating room of the future. *Annals of Biomedical Engineering*, 48(8):2171–2191, 2020.
- [16] Melissa Whitworth, Leanne Bricker, and Clare Mullan. Ultrasound for fetal assessment in early pregnancy. *Cochrane database of systematic reviews*, (7), 2015.
- [17] Hilmar H Bijma, Agnes van der Heide, and Hajo IJ Wildschut. Decision-making after ultrasound diagnosis of fetal abnormality. *Reproductive health matters*, 16(sup31):82–89, 2008.
- [18] Qinghua Huang and Zhaozheng Zeng. A review on real-time 3d ultrasound imaging technology. *BioMed research international*, 2017, 2017.
- [19] Lu Meng, Dan Zhao, Zeyu Yang, and Bing Wang. Automatic display of fetal brain planes and automatic measurements of fetal brain parameters by

- transabdominal three-dimensional ultrasound. *Journal of Clinical Ultrasound*, 48(2):82–88, 2020.
- [20] Zeynettin Akkus, Jason Cai, Arunnit Boonrod, Atefeh Zeinoddini, Alexander D Weston, Kenneth A Philbrick, and Bradley J Erickson. A survey of deep-learning applications in ultrasound: Artificial intelligence–powered ultrasound for improving clinical workflow. *Journal of the American College of Radiology*, 16(9):1318–1328, 2019.
- [21] Abdeldjalil Ouahabi and Abdelmalik Taleb-Ahmed. Deep learning for real-time semantic segmentation: Application in ultrasound imaging. *Pattern Recognition Letters*, 144:27–34, 2021.
- [22] Yu-Ting Shen, Liang Chen, Wen-Wen Yue, and Hui-Xiong Xu. Artificial intelligence in ultrasound. *European Journal of Radiology*, 139:109717, 2021.
- [23] Jordina Torrents-Barrena, Gemma Piella, Narcís Masoller, Eduard Gratacós, Elisenda Eixarch, Mario Ceresa, and Miguel Ángel González Ballester. Segmentation and classification in mri and us fetal imaging: recent trends and future prospects. *Medical Image Analysis*, 51:61–88, 2019.
- [24] Chunlin Song, Tao Gao, Hong Wang, Sud Sudirman, Wei Zhang, and Haogang Zhu. The classification and segmentation of fetal anatomies ultrasound image: A survey. *Journal of Medical Imaging and Health Informatics*, 11(3):789–802, 2021.
- [25] S Jayanthi Sree and C Vasanthanayaki. Ultrasound fetal image segmentation techniques: a review. *Current Medical Imaging*, 15(1):52–60, 2019.
- [26] Patricia Garcia-Canadilla, Sergio Sanchez-Martinez, Fatima Crispi, and Bart Bijmens. Machine learning in fetal cardiology: what to expect. *Fetal diagnosis and therapy*, 47(5):363–372, 2020.
- [27] Shaine A Morris and Keila N Lopez. Deep learning for detecting congenital heart disease in the fetus. *Nature Medicine*, 27(5):764–765, 2021.
- [28] Vidhi Rawat, Alok Jain, and Vibhakar Shrimali. Automated techniques for the interpretation of fetal abnormalities: a review. *Applied bionics and biomechanics*, 2018, 2018.
- [29] S Nikkath Bushra and G Shobana. Obstetrics and gynaecology ultrasound image analysis towards cryptic pregnancy using deep learning-a review. In *2021 5th International Conference on Intelligent Computing and Control Systems (ICICCS)*, pages 949–953. IEEE, 2021.
- [30] Zhiyi Chen, Zhenyu Liu, Meng Du, and Ziyao Wang. Artificial intelligence in obstetric ultrasound: an update and future applications. *Frontiers in Medicine*, page 1431, 2021.

Bibliography

- [31] Pedro HB Diniz, Yi Yin, and Sally Collins. Deep learning strategies for ultrasound in pregnancy. *European Medical Journal. Reproductive health*, 6(1):73, 2020.
- [32] Rosana El Jurdi, Caroline Petitjean, Paul Honeine, Veronika Cheplygina, and Fahed Abdallah. High-level prior-based loss functions for medical image segmentation: A survey. *Computer Vision and Image Understanding*, 210:103248, 2021.
- [33] Qi Wang, Yue Ma, Kun Zhao, and Yingjie Tian. A comprehensive survey of loss functions in machine learning. *Annals of Data Science*, 9(2):187–212, 2022.
- [34] Helen R Sofaer, Jennifer A Hoeting, and Catherine S Jarneveich. The area under the precision-recall curve as a performance metric for rare binary events. *Methods in Ecology and Evolution*, 10(4):565–577, 2019.
- [35] Orhun Utku Aydin, Abdel Aziz Taha, Adam Hilbert, Ahmed A Khalil, Ivana Galinovic, Jochen B Fiebach, Dietmar Frey, and Vince Istvan Madai. On the usage of average hausdorff distance for segmentation performance assessment: hidden error when used for ranking. *European Radiology Experimental*, 5(1):1–7, 2021.
- [36] Tilke Judd, Krista Ehinger, Frédo Durand, and Antonio Torralba. Learning to predict where humans look. In *2009 IEEE 12th International Conference on Computer Vision*, pages 2106–2113. IEEE, 2009.
- [37] Zoya Bylinskii, Tilke Judd, Aude Oliva, Antonio Torralba, and Frédo Durand. What do different evaluation metrics tell us about saliency models? *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(3):740–757, 2018.
- [38] Robert J Peters, Asha Iyer, Laurent Itti, and Christof Koch. Components of bottom-up gaze allocation in natural images. *Vision Research*, 45(18):2397–2416, 2005.
- [39] Xavier P Burgos-Artizzu, David Coronado-Gutiérrez, Brenda Valenzuela-Alcaraz, Elisenda Bonet-Carne, Elisenda Eixarch, Fatima Crispi, and Eduard Gratacós. Evaluation of deep convolutional neural networks for automatic classification of common maternal fetal ultrasound planes. *Scientific Reports*, 10(1):1–12, 2020.
- [40] Sylvia Rueda, Sana Fathima, Caroline L Knight, Mohammad Yaqub, Aris T Papageorgiou, Bahbib Rahmatullah, Alessandro Foi, Matteo Maggioni, Antonietta Pepe, Jussi Tohka, et al. Evaluation and comparison of current fetal ultrasound image segmentation methods for biometric measurements: a grand challenge. *IEEE Transactions on Medical Imaging*, 33(4):797–813, 2013.

- [41] Thomas LA van den Heuvel, Dagmar de Bruijn, Chris L de Korte, and Bram van Ginneken. Automated measurement of fetal head circumference using 2D ultrasound images. *PloS One*, 13(8):e0200412, 2018.
- [42] Lingyun Wu, Jie-Zhi Cheng, Shengli Li, Baiying Lei, Tianfu Wang, and Dong Ni. Fuiqa: Fetal ultrasound image quality assessment with deep convolutional networks. *IEEE Transactions on Cybernetics*, 47(5):1336–1349, 2017.
- [43] Zhen Yu, Ee-Leng Tan, Dong Ni, Jing Qin, Siping Chen, Shengli Li, Baiying Lei, and Tianfu Wang. A deep convolutional neural network-based framework for automatic fetal facial standard plane recognition. *IEEE Journal of Biomedical and Health Informatics*, 22(3):874–885, 2017.
- [44] Ruowei Qu, Guizhi Xu, Chunxia Ding, Wenyan Jia, and Mingui Sun. Standard plane identification in fetal brain ultrasound scans using a differential convolutional neural network. *IEEE Access*, 8:83821–83830, 2020.
- [45] Peiyao Kong, Dong Ni, Siping Chen, Shengli Li, Tianfu Wang, and Baiying Lei. Automatic and efficient standard plane recognition in fetal ultrasound images via multi-scale dense networks. In *Data Driven Treatment Response Assessment and Preterm, Perinatal, and Paediatric Image Analysis*, pages 160–168. Springer, 2018.
- [46] Jiajun Liang, Rian Huang, Peiyao Kong, Shengli Li, Tianfu Wang, and Baiying Lei. SPRNet: Automatic fetal standard plane recognition network for ultrasound images. In *Smart Ultrasound Imaging and Perinatal, Preterm and Paediatric Image Analysis*, pages 38–46. Springer, 2019.
- [47] Vaanathi Sundaresan, Christopher P Bridge, Christos Ioannou, and J Alison Noble. Automated characterization of the fetal heart in ultrasound images using fully convolutional neural networks. In *2017 IEEE 14th International Symposium on Biomedical Imaging*, pages 671–674. IEEE, 2017.
- [48] Alberto Montero, Elisenda Bonet-Carne, and Xavier Paolo Burgos-Artizzu. Generative adversarial networks to improve fetal brain fine-grained plane classification. *Sensors*, 21(23):7975, 2021.
- [49] Shuangchi He, Zehui Lin, Xin Yang, Chaoyu Chen, Jian Wang, Xue Shuang, Ziwei Deng, Qin Liu, Yan Cao, Xiduo Lu, et al. Statistical dependency guided contrastive learning for multiple labeling in prenatal ultrasound. In *International Workshop on Machine Learning in Medical Imaging*, pages 190–198. Springer, 2021.
- [50] Qingjie Meng, Daniel Rueckert, and Bernhard Kainz. Unsupervised cross-domain image classification by distance metric guided feature alignment. In

Bibliography

- Medical Ultrasound, and Preterm, Perinatal and Paediatric Image Analysis*, pages 146–157. Springer, 2020.
- [51] Hao Chen, Lingyun Wu, Qi Dou, Jing Qin, Shengli Li, Jie-Zhi Cheng, Dong Ni, and Pheng-Ann Heng. Ultrasound standard plane detection using a composite neural network framework. *IEEE Transactions on Cybernetics*, 47(6):1576–1586, 2017.
- [52] Bin Pu, Kenli Li, Shengli Li, and Ningbo Zhu. Automatic fetal ultrasound standard plane recognition based on deep learning and IIoT. *IEEE Transactions on Industrial Informatics*, 2021.
- [53] Jo Schlemper, Ozan Oktay, Michiel Schaap, Mattias Heinrich, Bernhard Kainz, Ben Glocker, and Daniel Rueckert. Attention gated networks: Learning to leverage salient regions in medical images. *Medical Image Analysis*, 53:197–207, 2019.
- [54] Yifan Cai, Harshita Sharma, Pierre Chatelain, and J Alison Noble. Multi-task sonoeenet: detection of fetal standardized planes assisted by generated sonographer attention maps. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 871–879. Springer, 2018.
- [55] Yifan Cai, Richard Droste, Harshita Sharma, Pierre Chatelain, Lior Drukker, Aris T Papageorghiou, and J Alison Noble. Spatio-temporal visual attention modelling of standard biometry plane-finding navigation. *Medical Image Analysis*, 65:101762, 2020.
- [56] Lok Hin Lee, Yuan Gao, and J Alison Noble. Principled ultrasound data augmentation for classification of standard planes. In *International Conference on Information Processing in Medical Imaging*, pages 729–741. Springer, 2021.
- [57] Jinbao Dong, Shengfeng Liu, Yimei Liao, Huaxuan Wen, Baiying Lei, Shengli Li, and Tianfu Wang. A generic quality control framework for fetal ultrasound cardiac four-chamber planes. *IEEE Journal of Biomedical and Health Informatics*, 24(4):931–942, 2019.
- [58] Zehui Lin, Minh Hung Le, Dong Ni, Siping Chen, Shengli Li, Tianfu Wang, and Baiying Lei. Quality assessment of fetal head ultrasound images based on faster R-CNN. In *Simulation, Image Processing, and Ultrasound Systems for Assisted Diagnosis and Navigation*, pages 38–46. Springer, 2018.
- [59] Zehui Lin, Shengli Li, Dong Ni, Yimei Liao, Huaxuan Wen, Jie Du, Siping Chen, Tianfu Wang, and Baiying Lei. Multi-task learning for quality assessment of fetal head ultrasound images. *Medical Image Analysis*, 58:101548, 2019.

- [60] Bo Zhang, Han Liu, Hong Luo, and Kejun Li. Automatic quality assessment for 2D fetal sonographic standard plane based on multitask learning. *Medicine*, 100(4), 2021.
- [61] Christian F Baumgartner, Konstantinos Kamnitsas, Jacqueline Matthew, Tara P Fletcher, Sandra Smith, Lisa M Koch, Bernhard Kainz, and Daniel Rueckert. SonoNet: real-time detection and localisation of fetal standard scan planes in freehand ultrasound. *IEEE Transactions on Medical Imaging*, 36(11):2204–2215, 2017.
- [62] Mohammad Yaqub, Brenda Kelly, Aris T Papageorghiou, and J Alison Noble. A deep learning solution for automatic fetal neurosonographic diagnostic plane verification using clinical standard constraints. *Ultrasound in Medicine & Biology*, 43(12):2925–2933, 2017.
- [63] Yuan Gao, Sridevi Beriwal, Rachel Craik, Aris T Papageorghiou, and J Alison Noble. Label efficient localization of fetal brain biometry planes in ultrasound through metric learning. In *Medical Ultrasound, and Preterm, Perinatal and Paediatric Image Analysis*, pages 126–135. Springer, 2020.
- [64] Liang Chen, Paul Bentley, Kensaku Mori, Kazunari Misawa, Michitaka Fujiwara, and Daniel Rueckert. Self-supervised learning for medical image analysis using image context restoration. *Medical Image Analysis*, 58:101539, 2019.
- [65] Jeremy Tan, Anselm Au, Qingjie Meng, and Bernhard Kainz. Semi-supervised learning of fetal anatomy from ultrasound. In *Domain Adaptation and Representation Transfer and Medical Image Learning with Less Labels and Imperfect Data*, pages 157–164. Springer, 2019.
- [66] Haoran Dou, Xin Yang, Jikuan Qian, Wufeng Xue, Hao Qin, Xu Wang, Lequan Yu, Shujun Wang, Yi Xiong, Pheng-Ann Heng, et al. Agent with warm start and active termination for plane localization in 3D ultrasound. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 290–298. Springer, 2019.
- [67] Xin Yang, Haoran Dou, Ruobing Huang, Wufeng Xue, Yuhao Huang, Jikuan Qian, Yuanji Zhang, Huanjia Luo, Huizhi Guo, Tianfu Wang, et al. Agent with warm start and adaptive dynamic termination for plane localization in 3D ultrasound. *IEEE Transactions on Medical Imaging*, 2021.
- [68] Xin Yang, Yuhao Huang, Ruobing Huang, Haoran Dou, Rui Li, Jikuan Qian, Xiaoqiong Huang, Wenlong Shi, Chaoyu Chen, Yuanji Zhang, et al. Searching collaborative agents for multi-plane localization in 3D ultrasound. *Medical Image Analysis*, page 102119, 2021.

Bibliography

- [69] Yuanwei Li, Bishesh Khanal, Benjamin Hou, Amir Alansary, Juan J Cerrolaza, Matthew Sinclair, Jacqueline Matthew, Chandni Gupta, Caroline Knight, Bernhard Kainz, et al. Standard plane detection in 3D fetal ultrasound using an iterative transformation network. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 392–400. Springer, 2018.
- [70] Pei-Yin Tsai, Ching-Hui Hung, Chi-Yeh Chen, and Yung-Nien Sun. Automatic fetal middle sagittal plane detection in ultrasound using generative adversarial network. *Diagnostics*, 11(1):21, 2021.
- [71] LJ Salomon, Z Alfirevic, F Da Silva Costa, RL Deter, F Figueras, Tet al Ghi, P Glanc, A Khalil, W Lee, R Napolitano, et al. ISUOG practice guidelines: ultrasound assessment of fetal biometry and growth. *Ultrasound in Obstetrics & Gynecology*, 53(6):715–723, 2019.
- [72] Dong Ni, Xin Yang, Xin Chen, Chien-Ting Chin, Siping Chen, Pheng Ann Heng, Shengli Li, Jing Qin, and Tianfu Wang. Standard plane localization in ultrasound by radial component model and selective search. *Ultrasound in Medicine & Biology*, 40(11):2728–2742, 2014.
- [73] Jinbao Dong, Shengfeng Liu, and Tianfu Wang. ARVBNNet: Real-time detection of anatomical structures in fetal ultrasound cardiac four-chamber planes. In *Machine Learning and Medical Engineering for Cardiovascular Health and Intravascular Imaging and Computer Assisted Stenting*, pages 130–137. Springer, 2019.
- [74] Arijit Patra and J Alison Noble. Multi-anatomy localization in fetal echocardiography videos. In *2019 IEEE 16th International Symposium on Biomedical Imaging*, pages 1761–1764. IEEE, 2019.
- [75] Weilin Huang, Christopher P Bridge, J Alison Noble, and Andrew Zisserman. Temporal HeartNet: towards human-level automatic analysis of fetal cardiac screening video. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 341–349. Springer, 2017.
- [76] Arijit Patra, Weilin Huang, and J Alison Noble. Learning spatio-temporal aggregation for fetal heart analysis in ultrasound video. In *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*, pages 276–284. Springer, 2017.
- [77] Yuan Gao and J Alison Noble. Detection and characterization of the fetal heartbeat in free-hand ultrasound sweeps with weakly-supervised two-streams convolutional networks. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 305–313. Springer, 2017.

- [78] MN Rachmatullah, Siti Nurmaini, AI Sapitri, A Darmawahyuni, B Tutuko, and Firdaus Firdaus. Convolutional neural network for semantic segmentation of fetal echocardiography based on four-chamber view. *Bulletin of Electrical Engineering and Informatics*, 10(4):1987–1996, 2021.
- [79] Lu Xu, Mingyuan Liu, Zhenrong Shen, Hua Wang, Xiaowei Liu, Xin Wang, Siyu Wang, Tiefeng Li, Shaomei Yu, Min Hou, et al. DW-Net: A cascaded convolutional neural network for apical four-chamber view segmentation in fetal echocardiography. *Computerized Medical Imaging and Graphics*, 80:101690, 2020.
- [80] Lu Xu, Mingyuan Liu, Jicong Zhang, and Yihua He. Convolutional-neural-network-based approach for segmentation of apical four-chamber view from fetal echocardiography. *IEEE Access*, 8:80437–80446, 2020.
- [81] Li Yu, Yi Guo, Yuanyuan Wang, Jinhua Yu, and Ping Chen. Segmentation of fetal left ventricle in echocardiographic sequences based on dynamic convolutional neural networks. *IEEE Transactions on Biomedical Engineering*, 64(8):1886–1895, 2017.
- [82] Shan An, Haogang Zhu, Yuanshuai Wang, Fangru Zhou, Xiaoxue Zhou, Xu Yang, Yingying Zhang, Xiangyu Liu, Zhicheng Jiao, and Yihua He. A category attention instance segmentation network for four cardiac chambers segmentation in fetal echocardiography. *Computerized Medical Imaging and Graphics*, 93:101983, 2021.
- [83] Siti Nurmaini, Muhammad Naufal Rachmatullah, Ade Iriani Sapitri, Annisa Darmawahyuni, Adithia Jovandy, Firdaus Firdaus, Bambang Tutuko, and Rossi Passarella. Accurate detection of septal defects with fetal ultrasonography images using deep learning-based multiclass instance segmentation. *IEEE Access*, 8:196160–196174, 2020.
- [84] Siti Nurmaini, Muhammad Naufal Rachmatullah, Ade Iriani Sapitri, Annisa Darmawahyuni, Bambang Tutuko, Firdaus Firdaus, Radiyati Umi Partan, and Nuswil Bernolian. Deep learning-based computer-aided fetal echocardiography: Application to heart standard view segmentation for congenital heart defects detection. *Sensors*, 21(23):8007, 2021.
- [85] Jeremy Tan, Anselm Au, Qingjie Meng, Sandy FinesilverSmith, John Simpson, Daniel Rueckert, Reza Razavi, Thomas Day, David Lloyd, and Bernhard Kainz. Automated detection of congenital heart disease in fetal ultrasound screening. In *Medical Ultrasound, and Preterm, Perinatal and Paediatric Image Analysis*, pages 243–252. Springer, 2020.
- [86] Ai Dozen, Masaaki Komatsu, Akira Sakai, Reina Komatsu, Kanto Shozu, Hide-nori Machino, Suguru Yasutomi, Tatsuya Arakaki, Ken Asada, Syuzo Kaneko,

Bibliography

- et al. Image segmentation of the ventricular septum in fetal cardiac ultrasound videos based on deep learning using time-series information. *Biomolecules*, 10(11):1526, 2020.
- [87] Masaaki Komatsu, Akira Sakai, Reina Komatsu, Ryu Matsuoka, Suguru Yasutomi, Kanto Shozu, Ai Dozen, Hidenori Machino, Hirokazu Hidaka, Tatsuya Arakaki, et al. Detection of cardiac structural abnormalities in fetal ultrasound videos using deep learning. *Applied Sciences*, 11(1):371, 2021.
- [88] Rima Arnaout, Lara Curran, Yili Zhao, Jami C Levine, Erin Chinn, and Anita J Moon-Grady. An ensemble of neural networks provides expert-level prenatal detection of complex congenital heart disease. *Nature Medicine*, 27(5):882–891, 2021.
- [89] Sibao Qiao, Shanchen Pang, Gang Luo, Silin Pan, Zengchen Yu, Taotao Chen, and Zhihan Lv. Rlds: An explainable residual learning diagnosis system for fetal congenital heart disease. *Future Generation Computer Systems*, 128:205–218, 2022.
- [90] Yuxin Gong, Yingying Zhang, Haogang Zhu, Jing Lv, Qian Cheng, Hongjia Zhang, Yihua He, and Shuliang Wang. Fetal congenital heart disease echocardiogram screening based on DGACNN: adversarial one-class classification combined with video transfer learning. *IEEE Transactions on Medical Imaging*, 39(4):1206–1222, 2019.
- [91] Bin Pu, Ningbo Zhu, Kenli Li, and Shengli Li. Fetal cardiac cycle detection in multi-resource echocardiograms using hybrid classification framework. *Future Generation Computer Systems*, 115:825–836, 2021.
- [92] Arijit Patra and Julia Alison Noble. Hierarchical class incremental learning of anatomical structures in fetal echocardiography videos. *IEEE Journal of Biomedical and Health Informatics*, 24(4):1046–1058, 2020.
- [93] Shuo Wang, Yang Hua, Yunyun Cao, Tao Song, Zhengui Xue, Xiaoping Gong, Guanjie Wang, Ruhui Ma, and Haibing Guan. Deep learning based fetal middle cerebral artery segmentation in large-scale ultrasound images. In *2018 IEEE International Conference on Bioinformatics and Biomedicine*, pages 532–539, 2018.
- [94] Yuzhou Wu, Kuifang Shen, Zhigang Chen, and Jia Wu. Automatic measurement of fetal cavum septum pellucidum from ultrasound images using deep attention network. In *2020 IEEE International Conference on Image Processing*, pages 2511–2515, 2020.
- [95] Vishal Singh, Pradeeba Sridar, Jinman Kim, Ralph Nanan, N. Poornima, Shanmuga Priya, G. Sameera Reddy, Sathyabama Chandrasekaran, and Ramarathnam Krishnakumar. Semantic segmentation of cerebellum in 2D fetal

- ultrasound brain images using convolutional neural networks. *IEEE Access*, 9:85864–85873, 2021.
- [96] Lun Zhang, Junhua Zhang, Zonggui Li, and Yingchao Song. A multiple-channel and atrous convolution network for ultrasound image segmentation. *Medical Physics*, 47(12):6270–6285, 2020.
- [97] Ruobing Huang, Weidi Xie, and J Alison Noble. Vp-nets: Efficient automatic localization of key brain structures in 3D fetal neurosonography. *Medical Image Analysis*, 47:127–139, 2018.
- [98] Madeleine K. Wyburd, Mark Jenkinson, and Ana I. L. Namburete. Cortical plate segmentation using CNNs in 3D fetal ultrasound. In Bartłomiej W. Papież, Ana I. L. Namburete, Mohammad Yaqub, and J. Alison Noble, editors, *Medical Image Understanding and Analysis*, pages 56–68. Springer International Publishing, 2020.
- [99] Lorenzo Venturini, Aris T. Papageorghiou, J. Alison Noble, and Ana I. Namburete. Multi-task CNN for structural semantic segmentation in 3D fetal brain ultrasound. *Communications in Computer and Information Science*, page 164–173, 2020.
- [100] Linde S. Hesse and Ana I. L. Namburete. Improving U-Net segmentation with active contour based label correction. In *Annual Conference on Medical Image Understanding and Analysis*, 2020.
- [101] Xin Yang, Xu Wang, Yi Wang, Haoran Dou, Shengli Li, Huaxuan Wen, Yi Lin, Pheng-Ann Heng, and Dong Ni. Hybrid attention for automatic segmentation of whole fetal head in prenatal ultrasound volumes. *Computer Methods and Programs in Biomedicine*, 194:105519, 2020.
- [102] Ana I L Namburete, Weidi Xie, Mohammad Yaqub, Andrew Zisserman, and J Alison Noble. Fully-automated alignment of 3D fetal brain ultrasound to a canonical reference space using multi-task learning. *Medical Image Analysis*, 46:1–14, 2018.
- [103] Felipe Moser, Ruobing Huang, Aris T. Papageorghiou, Bartłomiej W. Papież, and Ana I. Namburete. Automated fetal brain extraction from clinical ultrasound volumes using 3D convolutional neural networks. *Communications in Computer and Information Science*, page 151–163, 2020.
- [104] Lok Hin Lee, Elizabeth Bradburn, Aris T Papageorghiou, and J Alison Noble. Calibrated bayesian neural networks to estimate gestational age and its uncertainty on fetal brain ultrasound images. In *Medical Ultrasound, and Preterm, Perinatal and Paediatric Image Analysis*, pages 13–22. Springer, 2020.

Bibliography

- [105] Ana IL Namburete, Weidi Xie, and J Alison Noble. Robust regression of brain maturation from 3D fetal neurosonography using CRNs. In *Fetal, Infant and Ophthalmic Medical Image Analysis*, pages 73–80. Springer, 2017.
- [106] Madeleine K Wyburd, Linde S Hesse, Moska Aliasi, Mark Jenkinson, Aris T Papageorghiou, Monique C Haak, and Ana IL Namburete. Assessment of regional cortical development through fissure based gestational age estimation in 3D fetal ultrasound. In *Uncertainty for Safe Utilization of Machine Learning in Medical Imaging, and Perinatal Imaging, Placental and Preterm Image Analysis*, pages 242–252. Springer, 2021.
- [107] HN Xie, Nan Wang, Miao He, LH Zhang, HM Cai, JB Xian, MF Lin, Jv Zheng, and YZ Yang. Using deep-learning algorithms to classify fetal brain ultrasound images as normal or abnormal. *Ultrasound in Obstetrics & Gynecology*, 56(4):579–587, 2020.
- [108] Ricky Hu, Rohit Singla, Ryan Yan, Chantal Mayer, and Robert N. Rohling. Automated placenta segmentation with a convolutional neural network weighted by acoustic shadow detection. In *2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pages 6718–6723, 2019.
- [109] Zoe Hu, Ricky Hu, Ryan Yan, Chantal Mayer, Robert N Rohling, and Rohit Singla. Automatic placenta abnormality detection using convolutional neural networks on ultrasound texture. In *Uncertainty for Safe Utilization of Machine Learning in Medical Imaging, and Perinatal Imaging, Placental and Preterm Image Analysis*, pages 147–156. Springer, 2021.
- [110] Veronika A Zimmer, Alberto Gomez, Emily Skelton, Nooshin Ghavami, Robert Wright, Lei Li, Jacqueline Matthew, Joseph V Hajnal, and Julia A Schnabel. A multi-task approach using positional information for ultrasound placenta segmentation. In *Medical Ultrasound, and Preterm, Perinatal and Paediatric Image Analysis*, pages 264–273. Springer, 2020.
- [111] Baris U Oguz, Jiancong Wang, Natalie Yushkevich, Alison Pouch, James Gee, Paul A Yushkevich, Nadav Schwartz, and Ipek Oguz. Combining deep learning and multi-atlas label fusion for automated placenta segmentation from 3DUS. In *Data Driven Treatment Response Assessment and Preterm, Perinatal, and Paediatric Image Analysis*, pages 138–148. Springer, 2018.
- [112] Jordina Torrents-Barrena, Gemma Piella, Brenda Valenzuela-Alcaraz, Eduard Gratacos, Elisenda Eixarch, Mario Ceresa, and Miguel A González Ballester. TTTS-STgan: Stacked generative adversarial networks for TTTS fetal surgery planning based on 3D ultrasound. *IEEE Transactions on Medical Imaging*, 39(11):3595–3606, 2020.

- [113] Pádraig Looney, Gordon N Stevenson, Kypros H Nicolaides, Walter Plasencia, Malid Molloholli, Stavros Natsis, and Sally L Collins. Fully automated, real-time 3D ultrasound segmentation to estimate first trimester placental volume using deep learning. *JCI insight*, 3(11), 2018.
- [114] Pádraig Looney, Gordon N Stevenson, Kypros H Nicolaides, Walter Plasencia, Malid Molloholli, Stavros Natsis, and Sally L Collins. Automatic 3D ultrasound segmentation of the first trimester placenta using deep learning. In *2017 IEEE 14th International Symposium on Biomedical Imaging*, pages 279–282. IEEE, 2017.
- [115] Veronika A Zimmer, Alberto Gomez, Emily Skelton, Nicolas Toussaint, Tong Zhang, Bishesh Khanal, Robert Wright, Yohan Noh, Alison Ho, Jacqueline Matthew, et al. Towards whole placenta segmentation at late gestation using multi-view ultrasound images. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 628–636. Springer, 2019.
- [116] Xin Yang, Lequan Yu, Shengli Li, Xu Wang, Na Wang, Jing Qin, Dong Ni, and Pheng-Ann Heng. Towards automatic semantic segmentation in volumetric ultrasound. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 711–719. Springer, 2017.
- [117] Nadav Schwartz, Ipek Oguz, Jiancong Wang, Alison Pouch, Natalie Yushkevich, Shobhana Parameshwaran, James Gee, Paul Yushkevich, and Baris Oguz. Fully automated placental volume quantification from 3DUS for prediction of small-for-gestational-age infants. *Journal of Ultrasound in Medicine*, 2021.
- [118] Jordina Torrents-Barrena, Núria Monill, Gemma Piella, Eduard Gratacós, Elisenda Eixarch, Mario Ceresa, and Miguel A González Ballester. Assessment of radiomics and deep learning for the segmentation of fetal and maternal anatomy in magnetic resonance imaging and ultrasound. *Academic Radiology*, 28(2):173–188, 2021.
- [119] Yan Li, Rong Xu, Jun Ohya, and Hiroyasu Iwata. Automatic fetal body and amniotic fluid segmentation from fetal ultrasound images by encoder-decoder network with inner layers. In *2017 39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pages 1485–1488. IEEE, 2017.
- [120] Hyun Cheol Cho, Siyu Sun, Chang Min Hyun, Ja-Young Kwon, Bukweon Kim, Yejin Park, and Jin Keun Seo. Automated ultrasound assessment of amniotic fluid index using deep learning. *Medical Image Analysis*, 69:101951, 2021.
- [121] Siyu Sun, Ja-Young Kwon, Yejin Park, Hyun Cheol Cho, Chang Min Hyun, and Jin Keun Seo. Complementary network for accurate amniotic fluid segmentation from ultrasound images. *IEEE Access*, 9:108223–108235, 2021.

Bibliography

- [122] Tai-Hui Xia, Man Tan, Jing-Hua Li, Jing-Jing Wang, Qing-Qing Wu, and De-Xing Kong. Establish a normal fetal lung gestational age grading model and explore the potential value of deep learning algorithms in fetal lung maturity evaluation. *Chinese Medical Journal*, 134(15):1828, 2021.
- [123] Ping Chen, Yunqi Chen, Yinhui Deng, Yuanyuan Wang, Ping He, Xiaoli Lv, and Jinhua Yu. A preliminary study to quantitatively evaluate the development of maturation degree for fetal lung based on transfer learning deep model from ultrasound images. *International Journal of Computer Assisted Radiology and Surgery*, 15(8):1407–1415, 2020.
- [124] Nipuna Weerasinghe, Nigel Hamilton Lovell, Alec William Welsh, and Gordon Stevenson. Multi-parametric fusion of 3D power doppler ultrasound for fetal kidney segmentation using fully convolutional neural networks. *IEEE Journal of Biomedical and Health Informatics*, 2020.
- [125] Astrid Franz, Alexander Schmidt-Richberg, Eliza Orasanu, and Cristian Lorenz. Deep learning-based spine centerline extraction in fetal ultrasound. In *Bildverarbeitung für die Medizin 2021*, pages 263–268. Springer, 2021.
- [126] Lei Chen, Yingying Tian, and Yujie Deng. Neural network algorithm-based three-dimensional ultrasound evaluation in the diagnosis of fetal spina bifida. *Scientific Programming*, 2021, 2021.
- [127] Alexander Schmidt-Richberg, Tom Brosch, Nicole Schadewaldt, Tobias Klinder, Angelo Cavallaro, Ibtisam Salim, David Roundhill, Aris Papageorgiou, and Cristian Lorenz. Abdomen segmentation in 3D fetal ultrasound using CNN-powered deformable models. In *Fetal, Infant and Ophthalmic Medical Image Analysis*, pages 52–61. Springer, 2017.
- [128] Richard Droste, Yifan Cai, Harshita Sharma, Pierre Chatelain, Aris T Papa-georgiou, and J Alison Noble. Towards capturing sonographic experience: cognition-inspired ultrasound video saliency prediction. In *Annual Conference on Medical Image Understanding and Analysis*, pages 174–186. Springer, 2019.
- [129] Xu Wang, Xin Yang, Haoran Dou, Shengli Li, Pheng-Ann Heng, and Dong Ni. Joint segmentation and landmark localization of fetal femur in ultrasound volumes. In *2019 IEEE EMBS International Conference on Biomedical & Health Informatics (BHI)*, pages 1–5. IEEE, 2019.
- [130] Juan J. Cerrolaza, Matthew Sinclair, Yuanwei Li, Alberto Gomez, Enzo Ferrante, Jaqueline Matthew, Chandni Gupta, Caroline L. Knight, and Daniel Rueckert. Deep learning with ultrasound physics for fetal skull segmentation. In *2018 IEEE 15th International Symposium on Biomedical Imaging*, pages 564–567, 2018.

- [131] Jorge Perez-Gonzalez, Nidiyare Hevia Montiel, and Verónica Medina Bañuelos. Deep learning spatial compounding from multiple fetal head ultrasound acquisitions. In *Medical Ultrasound, and Preterm, Perinatal and Paediatric Image Analysis*, pages 305–314. Springer, 2020.
- [132] Lu Wang, Dong Guo, Guotai Wang, and Shaoting Zhang. Annotation-efficient learning for medical image segmentation based on noisy pseudo labels and adversarial learning. *IEEE Transactions on Medical Imaging*, 2020.
- [133] Tejal Singh, Srinivas Rao Kudavelly, and K Venkata Suryanarayana. Deep learning based fetal face detection and visualization in prenatal ultrasound. In *2021 IEEE 18th International Symposium on Biomedical Imaging*, pages 1760–1763. IEEE, 2021.
- [134] Chaoyu Chen, Xin Yang, Ruobing Huang, Wenlong Shi, Shengfeng Liu, Mingrong Lin, Yuhao Huang, Yong Yang, Yuanji Zhang, Huanjia Luo, et al. Region proposal network with graph prior and IoU-balance loss for landmark detection in 3D ultrasound. In *2020 IEEE 17th International Symposium on Biomedical Imaging*, pages 1–5. IEEE, 2020.
- [135] Yasunari Miyagi, Toshiyuki Hata, Saori Bouno, Aya Koyanagi, and Takahito Miyake. Recognition of facial expression of fetuses by artificial intelligence (AI). *Journal of Perinatal Medicine*, 49(5):596–603, 2021.
- [136] Sagar Vaze and Ana IL Namburete. Segmentation of fetal adipose tissue using efficient cnns for portable ultrasound. In *Data Driven Treatment Response Assessment and Preterm, Perinatal, and Paediatric Image Analysis*, pages 55–65. Springer, 2018.
- [137] Fajar Astuti Hermawati, Handayani Tjandrasa, and Nanik Suciati. Phase-based thresholding schemes for segmentation of fetal thigh cross-sectional region in ultrasound images. *Journal of King Saud University-Computer and Information Sciences*, 2021.
- [138] Pradeeba Sridar, Ashnil Kumar, Ann Quinton, Ralph Nanan, Jinman Kim, and Ramarathnam Krishnakumar. Decision fusion-based fetal ultrasound image plane classification using convolutional neural networks. *Ultrasound in Medicine & Biology*, 45(5):1259–1273, 2019.
- [139] Genta Ishikawa, Rong Xu, Jun Ohya, and Hiroyasu Iwata. Detecting a fetus in ultrasound images using grad cam and locating the fetus in the uterus. In *International Conference on Pattern Recognition Applications and Methods*, pages 181–189, 2019.
- [140] Harshita Sharma, Richard Droste, Pierre Chatelain, Lior Drukker, Aris T Papageorghiou, and J Alison Noble. Spatio-temporal partitioning and descrip-

Bibliography

- tion of full-length routine fetal anomaly ultrasound scans. In *2019 IEEE 16th International Symposium on Biomedical Imaging*, pages 987–990. IEEE, 2019.
- [141] Mohammad Alsharid, Rasheed El-Bouri, Harshita Sharma, Lior Drukker, Aris T Papageorghiou, and J Alison Noble. A curriculum learning based approach to captioning ultrasound images. In *Medical Ultrasound, and Preterm, Perinatal and Paediatric Image Analysis*, pages 75–84. Springer, 2020.
- [142] Qingchao Chen, Yang Liu, Yipeng Hu, Alice Self, Aris Papageorghiou, and J Alison Noble. Cross-device cross-anatomy adaptation network for ultrasound video analysis. In *Medical Ultrasound, and Preterm, Perinatal and Paediatric Image Analysis*, pages 42–51. Springer, 2020.
- [143] Zhoubing Xu, Yuankai Huo, JinHyeong Park, Bennett Landman, Andy Milkowski, Sasa Grbic, and Shaohua Zhou. Less is more: Simultaneous view classification and landmark detection for abdominal ultrasound images. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 711–719. Springer, 2018.
- [144] Yuan Gao and J Alison Noble. Learning and understanding deep spatio-temporal representations from free-hand fetal ultrasound sweeps. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 299–308. Springer, 2019.
- [145] Lingyun Wu, Yang Xin, Shengli Li, Tianfu Wang, Pheng-Ann Heng, and Dong Ni. Cascaded fully convolutional networks for automatic prenatal ultrasound image segmentation. In *IEEE 14th International Symposium on Biomedical Imaging*, pages 663–666. IEEE, 2017.
- [146] Zhendong Liu, Xiaoqiong Huang, Xin Yang, Rui Gao, Rui Li, Yuanji Zhang, Yankai Huang, Guangquan Zhou, Yi Xiong, Alejandro F Frangi, et al. Generalize ultrasound image segmentation via instant and plug & play style transfer. In *2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI)*, pages 419–423. IEEE, 2021.
- [147] Xin Yang, Lequan Yu, Shengli Li, Huaxuan Wen, Dandan Luo, Cheng Bian, Jing Qin, Dong Ni, and Pheng-Ann Heng. Towards automated semantic segmentation in prenatal volumetric ultrasound. *IEEE Transactions on Medical Imaging*, 38:180–193, 2019.
- [148] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017.

- [149] LJ Salomon, Z Alfirevic, V Berghella, CM Bilardo, GE Chalouhi, F Da Silva Costa, E Hernandez-Andrade, G Malinger, H Munoz, D Paladini, et al. ISUOG practice guidelines (updated): performance of the routine mid-trimester fetal ultrasound scan. *Ultrasound in obstetrics & gynecology: the official journal of the International Society of Ultrasound in Obstetrics and Gynecology*, 2022.
- [150] Valerie A Luyckx and Barry M Brenner. Birth weight, malnutrition and kidney-associated outcomes—a global concern. *Nature Reviews Nephrology*, 11(3):135–149, 2015.
- [151] Janaína Campos Senra, Carlos Tadashi Yoshizaki, Giovana Farina Doro, Rodrigo Ruano, Maria Augusta Bento Cicaroni Gibelli, Agatha Sacramento Rodrigues, Vera Hermina Kalika Koch, Vera Lúcia Jornada Krebs, Marcelo Zugaib, Rossana Pulcineli Vieira Francisco, et al. Kidney impairment in fetal growth restriction: three-dimensional evaluation of volume and vascularization. *Prenatal Diagnosis*, 40(11):1408–1417, 2020.
- [152] Yunpeng Chen, Haoqi Fan, Bing Xu, Zhicheng Yan, Yannis Kalantidis, Marcus Rohrbach, Shuicheng Yan, and Jiashi Feng. Drop an octave: Reducing spatial redundancy in convolutional neural networks with octave convolution. In *IEEE/CVF International Conference on Computer Vision*, pages 3435–3444, 2019.
- [153] Matthew Sinclair, Christian F Baumgartner, Jacqueline Matthew, Wenjia Bai, Juan Cerrolaza Martinez, Yuanwei Li, Sandra Smith, Caroline L Knight, Bernhard Kainz, Jo Hajnal, et al. Human-level performance on automatic head biometrics in fetal ultrasound using fully convolutional neural networks. In *40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pages 714–717. IEEE, 2018.
- [154] Yibiao Rong, Dehui Xiang, Weifang Zhu, Fei Shi, Enting Gao, Zhun Fan, and Xinjian Chen. Deriving external forces via convolutional neural networks for biomedical image segmentation. *Biomedical Optics Express*, 10(8):3800–3814, 2019.
- [155] Everton Leonardo Skeika, Mathias Rodrigues Da Luz, Bruno José Torres Fernandes, Hugo Valadares Siqueira, and Mauren Louise Sguario Coelho De Andrade. Convolutional neural network to detect and measure fetal skull circumference in ultrasound imaging. *IEEE Access*, 8:191519–191529, 2020.
- [156] Yan Zeng, Po-Hsiang Tsui, Weiwei Wu, Zhu Huang Zhou, and Shuicai Wu. Fetal ultrasound image segmentation for automatic head circumference biometry using deeply supervised attention-gated V-Net. *Journal of Digital Imaging*, 34(1):134–148, 2021.

Bibliography

- [157] Cahya Perbawa Aji, Muhammad Hilman Fatoni, and Tri Arief Sardjono. Automatic measurement of fetal head circumference from 2-dimensional ultrasound. In *2019 International Conference on Computer Engineering, Network, and Intelligent Multimedia*, pages 1–5. IEEE, 2019.
- [158] Donghao Qiao and Farhana Zulkernine. Dilated squeeze-and-excitation U-Net for fetal ultrasound image segmentation. In *2020 IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology*, pages 1–7. IEEE, 2020.
- [159] Mostafa Ghelich Oghli, Shakiba Moradi, Nasim Sirjani, Reza Gerami, Payam Ghaderi, Ali Shabanzadeh, Hossein Arabi, Isaac Shiri, and Habib Zaidi. Automatic measurement of fetal head biometry from ultrasound images using deep neural networks. In *2020 IEEE Nuclear Science Symposium and Medical Imaging Conference*, pages 1–3. IEEE.
- [160] Prerna Bhalla, Ramesh Kumar Sunkaria, Aman Kamboj, and Anterpreet Kaur Bedi. Automatic fetus head segmentation in ultrasound images by attention based encoder decoder network. In *2021 12th International Conference on Computing Communication and Networking Technologies*, pages 1–7. IEEE, 2021.
- [161] Zahra Sobhaninia, Shima Rafiei, Ali Emami, Nader Karimi, Kayvan Najarian, Shadrokh Samavi, and SM Reza Soroushmehr. Fetal ultrasound image segmentation for measuring biometric parameters using multi-task deep learning. In *41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pages 6545–6548. IEEE, 2019.
- [162] Zahra Sobhaninia, Ali Emami, Nader Karimi, and Shadrokh Samavi. Localization of fetal head in ultrasound images by multiscale view and deep neural networks. In *2020 25th International Computer Conference, Computer Society of Iran*, pages 1–5. IEEE, 2020.
- [163] Peixuan Li, Huaici Zhao, Pengfei Liu, and Feidao Cao. Automated measurement network for accurate segmentation and parameter modification in fetal head ultrasound images. *Medical & Biological Engineering & Computing*, 58(11):2879–2892, 2020.
- [164] Hwa Pyung Kim, Sung Min Lee, Ja-Young Kwon, Yejin Park, Kang Cheol Kim, and Jin Keun Seo. Automatic evaluation of fetal head biometry from ultrasound images using machine learning. *Physiological Measurement*, 40(6):065009, 2019.
- [165] Baidaa Al-Bander, Theiab Alzahrani, Saeed Alzahrani, Bryan M Williams, and Yalin Zheng. Improving fetal head contour detection by object localisation with deep learning. In *Annual Conference on Medical Image Understanding and Analysis*, pages 142–150. Springer, 2019.

- [166] Maria Chiara Fiorentino, Sara Moccia, Morris Capparuccini, Sara Giamberini, and Emanuele Frontoni. A regression framework to head-circumference delineation from US fetal images. *Computer Methods and Programs in Biomedicine*, 198:105771, 2021.
- [167] Sara Moccia, Maria Chiara Fiorentino, and Emanuele Frontoni. Maskr²CNN: a distance-field regression version of Mask-RCNN for fetal-head delineation in ultrasound images. *International Journal of Computer Assisted Radiology and Surgery*, pages 1–8, 2021.
- [168] Yanda Meng, Meng Wei, Dongxu Gao, Yitian Zhao, Xiaoyun Yang, Xiaowei Huang, and Yalin Zheng. CNN-GCN aggregation enabled boundary regression for biomedical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 352–362. Springer, 2020.
- [169] Samuel Budd, Matthew Sinclair, Bishesh Khanal, Jacqueline Matthew, David Lloyd, Alberto Gomez, Nicolas Toussaint, Emma C Robinson, and Bernhard Kainz. Confident head circumference measurement from ultrasound with real-time feedback for sonographers. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 683–691. Springer, 2019.
- [170] Jing Zhang, Caroline Petitjean, Pierre Lopez, and Samia Ainouz. Direct estimation of fetal head circumference from ultrasound images based on regression CNN. In *Medical Imaging with Deep Learning*, pages 914–922. PMLR, 2020.
- [171] Jing Zhang, Caroline Petitjean, Florian Yger, and Samia Ainouz. Explainability for regression CNN in fetal head circumference estimation from ultrasound images. In *Interpretable and Annotation-Efficient Learning for Medical Image Computing*, pages 73–82. Springer, 2020.
- [172] Jaeseong Jang, Yejin Park, Bukweon Kim, Sung Min Lee, Ja-Young Kwon, and Jin Keun Seo. Automatic estimation of fetal abdominal circumference from ultrasound images. *IEEE Journal of Biomedical and Health Informatics*, 22(5):1512–1520, 2017.
- [173] Bukweon Kim, Kang Cheol Kim, Yejin Park, Ja-Young Kwon, Jaeseong Jang, and Jin Keun Seo. Machine-learning-based automatic identification of fetal abdominal circumference from ultrasound images. *Physiological measurement*, 39(10):105007, 2018.
- [174] Sevim Cengiz and Mohammad Yaqub. Automatic fetal gestational age estimation from first trimester scans. In *International Workshop on Advances in Simplifying Medical Ultrasound*, pages 220–227. Springer, 2021.

Bibliography

- [175] Xijie Chen, Miao He, Tingting Dan, Nan Wang, Meifang Lin, Lihe Zhang, Jianbo Xian, Hongmin Cai, and Hongning Xie. Automatic measurements of fetal lateral ventricles in 2d ultrasound images using deep learning. *Frontiers in Neurology*, 11:526, 2020.
- [176] Fengcheng Zhu, Mengyuan Liu, Feifei Wang, Di Qiu, Ruiman Li, and Chenyang Dai. Automatic measurement of fetal femur length in ultrasound images: a comparison of random forest regression model and SegNet. *Mathematical Biosciences and Engineering*, 18(6):7790–7805, 2021.
- [177] Jiancong Chen, Yingying Zhang, Jingyi Wang, Xiaoxue Zhou, Yihua He, and Tong Zhang. Ellipsenet: Anchor-free ellipse detection for automatic cardiac biometrics in fetal echocardiography. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 218–227. Springer, 2021.
- [178] Sophia Bano, Brian Dromey, Francisco Vasconcelos, Raffaele Napolitano, Anna L David, Donald M Peebles, and Danail Stoyanov. AutoFB: Automating fetal biometry estimation from standard ultrasound planes. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 228–238. Springer, 2021.
- [179] Szymon Płotka, Tomasz Włodarczyk, Adam Klasa, Michał Lipa, Arkadiusz Sitek, and Tomasz Trzciński. Fetalnet: Multi-task deep learning framework for fetal ultrasound biometric measurements. In *International Conference on Neural Information Processing*, pages 257–265. Springer, 2021.
- [180] Hosuk Ryou, Mohammad Yaqub, Angelo Cavallaro, Aris T Papageorghiou, and J Alison Noble. Automated 3D ultrasound image analysis for first trimester assessment of fetal health. *Physics in Medicine & Biology*, 64(18):185010, 2019.
- [181] Mostafa Ghelich Oghli, Ali Shabanzadeh, Shakiba Moradi, Nasim Sirjani, Reza Gerami, Payam Ghaderi, Morteza Sanei Taheri, Isaac Shiri, Hossein Arabi, and Habib Zaidi. Automatic fetal biometry prediction using a novel deep convolutional network architecture. *Physica Medica*, 88:127–137, 2021.
- [182] Yuan Gao, Lokhin Lee, Richard Droste, Rachel Craik, Sridevi Beriwal, Aris Papageorghiou, and Alison Noble. A dual adversarial calibration framework for automatic fetal brain biometry. In *IEEE/CVF International Conference on Computer Vision*, pages 3246–3254, 2021.
- [183] Juan C Prieto, Hina Shah, Alan J Rosenbaum, Xiaoning Jiang, Patrick Munsonda, Joan T Price, Elizabeth M Stringer, Bellington Vwalika, David M Stamilio, and Jeffrey SA Stringer. An automated framework for image classification and segmentation of fetal ultrasound images for gestational age

- estimation. In *Medical Imaging 2021: Image Processing*, volume 11596, page 115961N. International Society for Optics and Photonics, 2021.
- [184] Khalid Rasheed, Faraz Junejo, Ayesha Malik, and Muhammad Saqib. Automated fetal head classification and segmentation using ultrasound video. *IEEE Access*, 2021.
- [185] Simon Bohlender, Ilkay Oksuz, and Anirban Mukhopadhyay. A survey on shape-constraint deep learning for medical image segmentation. *IEEE Reviews in Biomedical Engineering*, 2021.
- [186] Alessandro Casella, Sara Moccia, Dario Paladini, Emanuele Frontoni, Elena De Momi, and Leonardo S Mattos. A shape-constraint adversarial framework with instance-normalized spatio-temporal features for inter-fetal membrane segmentation. *Medical Image Analysis*, 70:102008, 2021.
- [187] Juan J Cerrolaza, Mirella López Picazo, Ludovic Humbert, Yoshinobu Sato, Daniel Rueckert, Miguel Ángel González Ballester, and Marius George Linguraru. Computational anatomy for multi-organ analysis in medical imaging: A review. *Medical Image Analysis*, 56:44–67, 2019.
- [188] Adrian Bulat and Georgios Tzimiropoulos. Human pose estimation via convolutional part heatmap regression. In *European Conference on Computer Vision*, pages 717–732. Springer, 2016.
- [189] Praveen Paul Lakra, Ashish Kumar, Narayanaswamy Mohanram, Ganapathy Krishnamurthi, and Arun Kumar Thittai. Deep-learning based identification of frames containing foetal gender region during early second trimester ultrasound scanning. In *2019 IEEE International Ultrasonics Symposium*, pages 471–474. IEEE, 2019.
- [190] Xin Yang, Wenlong Shi, Haoran Dou, Jikuan Qian, Yi Wang, Wufeng Xue, Shengli Li, Dong Ni, and Pheng-Ann Heng. Fetusmap: Fetal pose estimation in 3D ultrasound. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 281–289. Springer, 2019.
- [191] Tomasz Włodarczyk, Szymon Płotka, Tomasz Trzciński, Przemysław Rokita, Nicole Sochacki-Wójcicka, Michał Lipa, and Jakub Wójcicki. Estimation of preterm birth markers with u-net segmentation network. *Smart Ultrasound Imaging and Perinatal, Preterm and Paediatric Image Analysis*, pages 95–103, 2019.
- [192] Tomasz Włodarczyk, Szymon Płotka, Przemysław Rokita, Nicole Sochacki-Wójcicka, Jakub Wójcicki, Michał Lipa, and Tomasz Trzciński. Spontaneous preterm birth prediction using convolutional neural networks. In *Medical*

Bibliography

- Ultrasound, and Preterm, Perinatal and Paediatric Image Analysis*, pages 274–283. Springer, 2020.
- [193] Yipeng Hu, Eli Gibson, Li-Lin Lee, Weidi Xie, Dean C Barratt, Tom Vercauteren, and J Alison Noble. Freehand ultrasound image simulation with spatially-conditioned generative adversarial networks. In *Molecular imaging, reconstruction and analysis of moving body organs, and stroke imaging and treatment*, pages 105–115. Springer, 2017.
- [194] Lin Zhang, Tiziano Portenier, Christoph Paulus, and Orcun Goksel. Deep image translation for enhancing simulated ultrasound images. In *Medical Ultrasound, and Preterm, Perinatal and Paediatric Image Analysis*, pages 85–94. Springer, 2020.
- [195] Yangdi Xu, Lok Hin Lee, Lior Drukker, Mohammad Yaqub, Aris T Papageorgiou, and Alison J Noble. Simulating realistic fetal neurosonography images with appearance and growth change using cycle-consistent adversarial networks and an evaluation. *Journal of Medical Imaging*, 7(5):057001, 2020.
- [196] Tianrui Liu, Qingjie Meng, Athanasios Vlontzos, Jeremy Tan, Daniel Rueckert, and Bernhard Kainz. Ultrasound video summarization using deep reinforcement learning. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 483–492. Springer, 2020.
- [197] Harshita Sharma, Lior Drukker, Pierre Chatelain, Richard Droste, Aris T Papageorgiou, and J Alison Noble. Knowledge representation and learning of operator clinical workflow from full-length routine fetal ultrasound scan videos. *Medical Image Analysis*, 69:101973, 2021.
- [198] Qingjie Meng, Christian Baumgartner, Matthew Sinclair, James Housden, Martin Rajchl, Alberto Gomez, Benjamin Hou, Nicolas Toussaint, Veronika Zimmer, Jeremy Tan, et al. Automatic shadow detection in 2D ultrasound images. In *Data Driven Treatment Response Assessment and Preterm, Perinatal, and Paediatric Image Analysis*, pages 66–75. Springer, 2018.
- [199] Qingjie Meng, Matthew Sinclair, Veronika Zimmer, Benjamin Hou, Martin Rajchl, Nicolas Toussaint, Ozan Oktay, Jo Schlemper, Alberto Gomez, James Housden, et al. Weakly supervised estimation of shadow confidence maps in fetal ultrasound imaging. *IEEE Transactions on Medical Imaging*, 38(12):2755–2767, 2019.
- [200] Qingjie Meng, Nick Pawlowski, Daniel Rueckert, and Bernhard Kainz. Representation disentanglement for multi-task learning with application to fetal ultrasound. In *Smart Ultrasound Imaging and Perinatal, Preterm and Paediatric Image Analysis*, pages 47–55. Springer, 2019.

- [201] Qingjie Meng, Jacqueline Matthew, Veronika A Zimmer, Alberto Gomez, David FA Lloyd, Daniel Rueckert, and Bernhard Kainz. Mutual information-based disentangled neural networks for classifying unseen categories in different domains: application to fetal ultrasound imaging. *IEEE Transactions on Medical Imaging*, 40(2):722–734, 2020.
- [202] Cheng Zhao, Richard Droste, Lior Drukker, Aris T Papageorgiou, and J Alison Noble. Visual-assisted probe movement guidance for obstetric ultrasound scanning using landmark retrieval. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 670–679. Springer, 2021.
- [203] Eberhard Merz and Sonila Pashaj. Advantages of 3d ultrasound in the assessment of fetal abnormalities. *Journal of perinatal medicine*, 45(6):643–650, 2017.
- [204] Farhan Mohamed and C Vei Siang. A survey on 3d ultrasound reconstruction techniques. *Artificial Intelligence—Applications in Medicine and Biology*, pages 73–92, 2019.
- [205] Mingyuan Luo, Xin Yang, Xiaoqiong Huang, Yuhao Huang, Yuxin Zou, Xindi Hu, Nishant Ravikumar, Alejandro F Frangi, and Dong Ni. Self context and shape prior for sensorless freehand 3D ultrasound reconstruction. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 201–210. Springer, 2021.
- [206] Micah J Sheller, Brandon Edwards, G Anthony Reina, Jason Martin, Sarthak Pati, Aikaterini Kotrotsou, Mikhail Milchenko, Weilin Xu, Daniel Marcus, Rivka R Colen, et al. Federated learning in medicine: facilitating multi-institutional collaborations without sharing patient data. *Scientific Reports*, 10(1):1–12, 2020.
- [207] Jacqueline Matthew, Emily Skelton, Thomas G Day, Veronika A Zimmer, Alberto Gomez, Gavin Wheeler, Nicolas Toussaint, Tianrui Liu, Samuel Budd, Karen Lloyd, et al. Exploring a new paradigm for the fetal anomaly ultrasound scan: Artificial intelligence in real time. *Prenatal Diagnosis*, 42(1):49–59, 2022.
- [208] S Degani. Fetal biometry: clinical, pathological, and technical considerations. *Obstetrical & Gynecological Survey*, 56(3):159–167, 2001.
- [209] Benjamin Deloison, Gihad E Chalouhi, Jean-Pierre Bernard, Yves Ville, and Laurent J Salomon. Outcomes of fetuses with small head circumference on second-trimester ultrasonography. *Prenatal Diagnosis*, 32(9):869–874, 2012.
- [210] I Putu Satwika, Ikhsanul Habibie, M Anwar Ma’sum, Andreas Febrian, and Enrico Budianto. Particle swarm optimization based 2-dimensional randomized

Bibliography

- Hough transform for fetal head biometry detection and approximation in ultrasound imaging. In *International Conference on Advanced Computer Science and Information System*, pages 468–473. IEEE, 2014.
- [211] Wei Lu and Jinglu Tan. Detection of incomplete ellipse in images with strong noise by iterative randomized Hough transform (IRHT). *Pattern Recognition*, 41(4):1268–1279, 2008.
- [212] Changming Sun. Automatic fetal head measurements from ultrasound images using circular shortest paths. *IEEE International Symposium on Biomedical Imaging, Challenge US: Biometric Measurements from Fetal Ultrasound Images*, pages 13–15, 2012.
- [213] Alessandro Foi, Matteo Maggioni, Antonietta Pepe, and Jussi Tohka. Head contour extraction from the fetal ultrasound images by difference of Gaussians revolved along elliptical paths. *IEEE International Symposium on Biomedical Imaging, Challenge US: Biometric Measurements from Fetal Ultrasound Images*, pages 1–3, 2012.
- [214] J L Perez-Gonzalez, J C Bello Muñoz, MC Rolon Porras, Fernando Arámbula-Cosío, and V Medina-Bañuelos. Automatic fetal head measurements from ultrasound images using optimal ellipse detection and texture maps. In *VI Latin American Congress on Biomedical Engineering*, pages 329–332. Springer, 2015.
- [215] V Rajinikanth, Nilanjan Dey, Rajesh Kumar, John Panneerselvam, and N Sri Madhava Raja. Fetal head periphery extraction from ultrasound image using Jaya algorithm and Chan-Vese segmentation. *Procedia Computer Science*, 152:66–73, 2019.
- [216] Richard V Stebbing and John E McManigle. A boundary fragment model for head segmentation in fetal ultrasound. *Challenge US: Biometric Measurements from Fetal Ultrasound Images*, pages 9–11, 2012.
- [217] Evelyn Arthur Anto, Benjamin Amoah, and Alessandro Crimi. Segmentation of ultrasound images of fetal anatomic structures using random forest for low-cost settings. In *37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pages 793–796. IEEE, 2015.
- [218] Gustavo Carneiro, Bogdan Georgescu, Sara Good, and Dorin Comaniciu. Detection and measurement of fetal anatomies from ultrasound images using a constrained probabilistic boosting tree. *IEEE Transactions on Medical Imaging*, 27(9):1342–1355, 2008.
- [219] Lei Zhang, Xujiong Ye, Tryphon Lambrou, Wenting Duan, Nigel Allinson, and Nicholas J Dudley. A supervised texton based approach for automatic

- segmentation and measurement of the fetal head and femur in 2D ultrasound images. *Physics in Medicine & Biology*, 61(3):1095, 2016.
- [220] Paolo Zafino, Guillaume Pernelle, Andre Mastmeyer, Alireza Mehrtash, Hongtao Zhang, Ron Kikinis, Tina Kapur, and Maria Francesca Spadea. Fully automatic catheter segmentation in MRI with 3D convolutional neural networks: application to MRI-guided gynecologic brachytherapy. *Physics in Medicine & Biology*, 64(16):165008, 2019.
- [221] Riccardo Rosati, Luca Romeo, Sonia Silvestri, Fabio Marcheggiani, Luca Tiano, and Emanuele Frontoni. Faster r-cnn approach for detection and quantification of dna damage in comet assay images. *Computers in Biology and Medicine*, page 103912, 2020.
- [222] Zahra Sobhaninia, Ali Emami, Nader Karimi, and Shadrokh Samavi. Localization of fetal head in ultrasound images by multiscale view and deep neural networks. In *International Computer Conference, Computer Society of Iran*, pages 1–5. IEEE, 2020.
- [223] Yanda Meng, Wei Meng, Dongxu Gao, Yitian Zhao, Xiaoyun Yang, Xiaowei Huang, and Yalin Zheng. Regression of instance boundary by aggregated CNN and GCN. In *European Conference on Computer Vision*, pages 190–207. Springer, 2020.
- [224] Xiaofei Du, Thomas Kurmann, Ping-Lin Chang, Maximilian Allan, Sebastien Ourselin, Raphael Sznitman, John D Kelly, and Danail Stoyanov. Articulated multi-instrument 2-D pose estimation using fully convolutional networks. *IEEE Transactions on Medical Imaging*, 37(5):1276–1287, 2018.
- [225] Emanuele Colleoni, Sara Moccia, Xiaofei Du, Elena De Momi, and Danail Stoyanov. Deep learning based robotic tool detection and articulation estimation with spatio-temporal layers. *IEEE Robotics and Automation Letters*, 4(3):2714–2721, 2019.
- [226] Georges Hattab, Marvin Arnold, Leon Strenger, Max Allan, Darja Arsentjeva, Oliver Gold, Tobias Simpfendörfer, Lena Maier-Hein, and Stefanie Speidel. Kidney edge detection in laparoscopic image data for computer-assisted surgery. *International Journal of Computer Assisted Radiology and Surgery*, pages 1–9, 2019.
- [227] Kezia Irene, Harlan Haidi, Nurul Faza, Winston Chandra, et al. Fetal head and abdomen measurement using convolutional neural network, Hough transform, and difference of Gaussian revolved along elliptical path (Dogell) algorithm. Available online at: *arXiv preprint arXiv:1911.06298*, 2019.

Bibliography

- [228] Joseph Redmon and Ali Farhadi. Yolo9000: better, faster, stronger. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 7263–7271, 2017.
- [229] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 234–241. Springer, 2015.
- [230] Nima Tajbakhsh, Laura Jeyaseelan, Qian Li, Jeffrey N Chiang, Zhihao Wu, and Xiaowei Ding. Embracing imperfect datasets: A review of deep learning solutions for medical image segmentation. *Medical Image Analysis*, page 101693, 2020.
- [231] Peter J Rousseeuw and Mia Hubert. Anomaly detection by robust statistics. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 8(2):e1236, 2018.
- [232] Dilip K Prasad, Maylor KH Leung, and Chai Quek. ElliFit: An unconstrained, non-iterative, least squares based geometric ellipse fitting method. *Pattern Recognition*, 46(5):1449–1465, 2013.
- [233] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask R-CNN. In *IEEE International Conference on Computer Vision*, pages 2961–2969, 2017.
- [234] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016.
- [235] Waleed Abdulla. Mask R-CNN for object detection and instance segmentation on keras and tensorflow. https://github.com/matterport/Mask_RCNN, 2017.
- [236] Yue Zhou, Houjin Chen, Yanfeng Li, Qin Liu, Xuanang Xu, Shu Wang, Pew-Thian Yap, and Dinggang Shen. Multi-task learning for segmentation and classification of tumors in 3d automated breast ultrasound images. *Medical Image Analysis*, 70:101918, 2021.
- [237] Xin Yi, Ekta Walia, and Paul Babyn. Generative adversarial network in medical imaging: A review. *Medical Image Analysis*, 58:101552, 2019.
- [238] Aldo Marzullo, Sara Moccia, Michele Catellani, Francesco Calimeri, and Elena De Momi. Towards realistic laparoscopic image generation using image-domain translation. *Computer Methods and Programs in Biomedicine*, page 105834, 2020.
- [239] Chenyu You, Ruihan Zhao, Lawrence H Staib, and James S Duncan. Momentum contrastive voxel-wise representation learning for semi-supervised volumetric

- medical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 639–652. Springer, 2022.
- [240] Phuc H Le-Khac, Graham Healy, and Alan F Smeaton. Contrastive representation learning: A framework and review. *IEEE Access*, 8:193907–193934, 2020.
- [241] Yipei Wang, Qianye Yang, Lior Drukker, Aris Papageorgiou, Yipeng Hu, and J Alison Noble. Task model-specific operator skill assessment in routine fetal ultrasound scanning. *International Journal of Computer Assisted Radiology and Surgery*, pages 1–8, 2022.
- [242] João Ramalhinho, Bongjin Koo, Nina Montaña-Brown, Shaheer U Saeed, Ester Bonmati, Kurinchi Gurusamy, Stephen P Pereira, Brian Davidson, Yipeng Hu, and Matthew J Clarkson. Deep hashing for global registration of untracked 2D laparoscopic ultrasound to CT. *International Journal of Computer Assisted Radiology and Surgery*, pages 1–8, 2022.
- [243] Gerda Bortsova, Florian Dubost, Laurens Hogeweg, Ioannis Katramados, and Marleen de Bruijne. Semi-supervised medical image segmentation via learning consistency under transformations. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 810–818. Springer, 2019.
- [244] Shekoofeh Azizi, Basil Mustafa, Fiona Ryan, Zachary Beaver, Jan Freyberg, Jonathan Deaton, Aaron Loh, Alan Karthikesalingam, Simon Kornblith, Ting Chen, et al. Big self-supervised models advance medical image classification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3478–3488, 2021.
- [245] Jianbo Jiao, Richard Droste, Lior Drukker, Aris T Papageorgiou, and J Alison Noble. Self-supervised representation learning for ultrasound video. In *2020 IEEE 17th International Symposium on Biomedical Imaging*, pages 1847–1850. IEEE, 2020.
- [246] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International Conference on Machine Learning*, pages 1597–1607. PMLR, 2020.
- [247] Yixiong Chen, Chunhui Zhang, Li Liu, Cheng Feng, Changfeng Dong, Yongfang Luo, and Xiang Wan. Uscl: pretraining deep ultrasound image diagnosis model through video contrastive representation learning. In *Medical Image Computing and Computer Assisted Intervention—MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part VIII 24*, pages 627–637. Springer, 2021.

Bibliography

- [248] Nima Tajbakhsh, Jae Y Shin, Suryakanth R Gurudu, R Todd Hurst, Christopher B Kendall, Michael B Gotway, and Jianming Liang. Convolutional neural networks for medical image analysis: Full training or fine tuning? *IEEE transactions on medical imaging*, 35(5):1299–1312, 2016.
- [249] Phillip M Cheng and Harshawn S Malhi. Transfer learning with convolutional neural networks for classification of abdominal ultrasound images. *Journal of digital imaging*, 30(2):234–243, 2017.
- [250] Sampa Misra, Seungwan Jeon, Ravi Managuli, Seiyon Lee, Gyuwon Kim, Chiho Yoon, Seungchul Lee, Richard G Barr, and Chulhong Kim. Bi-modal transfer learning for classifying breast cancers via combined b-mode and ultrasound strain imaging. *IEEE Transactions on Ultrasonics, Ferroelectrics, and Frequency Control*, 69(1):222–232, 2021.
- [251] Daniel F McWilliams and David A Walsh. Pain mechanisms in rheumatoid arthritis. *Clinical and Experimental Rheumatology*, 35(Suppl 107):94–101, 2017.
- [252] Daniel Aletaha, Julia Funovits, and Josef S Smolen. Physical disability in rheumatoid arthritis is associated with cartilage damage rather than bone destruction. *Annals of the rheumatic diseases*, 70(5):733–739, 2011.
- [253] Søren Torp-Pedersen, EM Bartels, J Wilhjelmsen, and H Bliddal. Articular cartilage thickness measured with us is not as easy as it appears: a systematic review of measurement techniques and image interpretation. *European Journal of Ultrasound*, 32(01):54–61, 2011.
- [254] Josef S Smolen, Désirée M van der Heijde, Edward C Keystone, Ronald F van Vollenhoven, Mary B Goldring, Benoît Guérette, Mary A Cifaldi, Naijun Chen, Shufang Liu, and Robert BM Landewé. Association of joint space narrowing with impairment of physical function and work ability in patients with early rheumatoid arthritis: protection beyond disease control by adalimumab plus methotrexate. *Annals of the rheumatic diseases*, 72(7):1156–1162, 2013.
- [255] Alexandra N Colebatch, Christopher John Edwards, Mikkel Østergaard, Désirée van der Heijde, Peter V Balint, Maria-Antonietta D’Agostino, Kristina Forslind, Walter Grassi, Espen A Haavardsholm, Glenn Haugeberg, et al. Eular recommendations for the use of imaging of the joints in the clinical management of rheumatoid arthritis. *Annals of the rheumatic diseases*, 72(6):804–814, 2013.
- [256] Uffe Møller Døhn, Bo J Ejbjerg, Maria Hasselquist, Eva Narvestad, Marcin Szkudlarek, Jakob M Møller, Henrik S Thomsen, Mikkel Østergaard, et al. Are bone erosions detected by magnetic resonance imaging and ultrasonography true erosions? a comparison with computed tomography in rheumatoid arthritis metacarpophalangeal joints. *Arthritis research & therapy*, 8(4):1–9, 2006.

- [257] Luana Mancarella, Olga Addimanda, Carlotta Cavallari, and Riccardo Meliconi. Synovial inflammation drives structural damage in hand osteoarthritis: a narrative literature review. *Current Rheumatology Reviews*, 13(1):43–50, 2017.
- [258] Win Min Oo, James M Linklater, Matthew Daniel, Simo Saarakkala, Jonathan Samuels, Philip G Conaghan, Helen I Keen, Leticia A Devezza, and David J Hunter. Clinimetrics of ultrasound pathologies in osteoarthritis: systematic literature review and meta-analysis. *Osteoarthritis and Cartilage*, 26(5):601–611, 2018.
- [259] Burkhard Möller, Harald Bonel, Marianne Rotzetter, Peter M Villiger, and Hans-rudolf Ziswiler. Measuring finger joint cartilage by ultrasound as a promising alternative to conventional radiograph imaging. *Arthritis Care & Research: Oficial Journal of the American College of Rheumatology*, 61(4):435– 441, 2009.
- [260] Edoardo Cipolletta, Emilio Filippucci, Andrea Di Matteo, Giulia Tesei, Micaela Ana Cosatti, Marco Di Carlo, and Walter Grassi. The reliability of ultrasound in the assessment of hyaline cartilage in rheumatoid arthritis and healthy metacarpal heads. *European Journal of Ultrasound*, 2020.
- [261] Edoardo Cipolletta, Jana Hurnakova, Andrea Di Matteo, Marco Di Carlo, Karel Pavelka, Walter Grassi, and Emilio Filippucci. Prevalence and distribution of cartilage and bone damage at metacarpal head in healthy subjects. *Clin Exp Rheumatol*, 2021.
- [262] Emilio Filippucci, Karine Rodrigues da Luz, Luca Di Geso, Fausto Salafi, Marika Tardella, Marina Carotti, Jamil Natour, and Walter Grassi. Inter-observer reliability of ultrasonography in the assessment of cartilage damage in rheumatoid arthritis. *Annals of the rheumatic diseases*, 69(10):1845–1848, 2010.
- [263] Peter Mandl, Paul Studenic, Emilio Filippucci, Artur Bachta, Marina Backhaus, David Bong, George AW Bruyn, Paz Collado, Nemanja Damjanov, Christian Dejaco, et al. Development of semiquantitative ultrasound scoring system to assess cartilage in rheumatoid arthritis. *Rheumatology*, 58(10):1802–1811, 2019.
- [264] P Mandl, G Supp, G Baksa, H Radner, P Studenic, J Gyebnar, R Kurucz, D Niedermayer, D Aletaha, P V Balint, et al. Relationship between radiographic joint space narrowing, sonographic cartilage thickness and anatomy in rheumatoid arthritis and control joints. *Annals of the Rheumatic Diseases*, 74(11):2022–2027, 2015.
- [265] Mohamed Mediouni, Daniel R. Schlatterer, Henning Madry, Magali Cucchiari, and Balwant Rai. A review of translational medicine. the future paradigm:

Bibliography

- how can we connect the orthopedic dots better? *Current medical research and opinion*, 34(7):1217–1229, 2018.
- [266] Tzen S Toh, Frank Dondelinger, and Dennis Wang. Looking beyond the hype: Applied ai and machine learning in translational medicine. *EBioMedicine*, 47:607–615, 2019.
- [267] Berend Stoel. Use of artificial intelligence in imaging in rheumatology—current status and future perspectives. *RMD open*, 6(1), 2020.
- [268] Gayatri Kompella, Maria Antico, Fumio Sasazawa, S Jeevakala, Keerthi Ram, Davide Fontanarosa, Ajay K Pandey, and Mohanasankar Sivaprakasam. Segmentation of femoral cartilage from knee ultrasound images using mask R-CNN. In *41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pages 966–969. IEEE, 2019.
- [269] Maria Antico, Fumio Sasazawa, Yu Takeda, ANJALI TUMKUR Jaiprakash, M-L Wille, A K Pandey, Ross Crawford, G Carneiro, and Davide Fontanarosa. Bayesian CNN for segmentation uncertainty inference on 4d ultrasound images of the femoral cartilage for guidance in robotic knee arthroscopy. *IEEE Access*, 2020.
- [270] Matteo Dunnhofer, Maria Antico, Fumio Sasazawa, Yu Takeda, Saskia Camps, Niki Martinel, Christian Micheloni, Gustavo Carneiro, and Davide Fontanarosa. Siam-U-Net: encoder-decoder siamese network for knee cartilage tracking in ultrasound images. *Medical Image Analysis*, 60:101631, 2020.
- [271] Md Belayet Hossain, Khin Wee Lai, Belinda Pingguan-Murphy, Yan Chai Hum, Maheza Irna Mohd Salim, and Yih Miin Liew. Contrast enhancement of ultrasound imaging of the knee joint cartilage for early detection of knee osteoarthritis. *Biomedical Signal Processing and Control*, 13:157–167, 2014.
- [272] Amir Faisal, Siew-Cheok Ng, Siew-Li Goh, and Khin Wee Lai. Knee cartilage segmentation and thickness computation from ultrasound images. *Medical & Biological Engineering & Computing*, 56(4):657–669, 2018.
- [273] Qinghua Huang, Fan Zhang, and Xuelong Li. Machine learning in ultrasound computer-aided diagnostic systems: a survey. *BioMed Research International*, 2018, 2018.
- [274] Prajna Ramesh Desai and Ilker Hacihaliloglu. Enhancement and automated segmentation of ultrasound knee cartilage for early diagnosis of knee osteoarthritis. In *IEEE 15th International Symposium on Biomedical Imaging*, pages 1471–1474. IEEE, 2018.
- [275] Prajna Desai and Ilker Hacihaliloglu. Knee-cartilage segmentation and thickness measurement from 2D ultrasound. *Journal of Imaging*, 5(4):43, 2019.

- [276] Georges Hattab, Marvin Arnold, Leon Strenger, Max Allan, Darja Arsentjeva, Oliver Gold, Tobias Simpfendörfer, Lena Maier-Hein, and Stefanie Speidel. Kidney edge detection in laparoscopic image data for computer-assisted surgery. *International Journal of Computer Assisted Radiology and Surgery*, 15(3):379–387, 2020.
- [277] Daniel Aletaha, Tuhina Neogi, Alan J Silman, Julia Funovits, David T Felson, Clifton O Bingham III, Neal S Birnbaum, Gerd R Burmester, Vivian P Bykerk, Marc D Cohen, et al. 2010 rheumatoid arthritis classification criteria: an american college of rheumatology/european league against rheumatism collaborative initiative. *Arthritis & rheumatism*, 62(9):2569–2581, 2010.
- [278] Kentaro Wada. labelme: Image Polygonal Annotation with Python. <https://github.com/wkentaro/labelme>, 2016.
- [279] Shi Yin, Qinmu Peng, Hongming Li, Zhengqiang Zhang, Xinge You, Katherine Fischer, Susan L Furth, Gregory E Tasian, and Yong Fan. Automatic kidney segmentation in ultrasound images using subsequent boundary distance regression and pixelwise classification networks. *Medical Image Analysis*, 60:101602, 2020.
- [280] Tongjie Y Zhang and Ching Y. Suen. A fast parallel algorithm for thinning digital patterns. *Communications of the ACM*, 27(3):236–239, 1984.
- [281] Ingrid Möller, Iustina Janta, Marina Backhaus, Sarah Ohrndorf, David A Bong, Carlo Martinoli, Emilio Filippucci, Luca Maria Sconfienza, Lene Terslev, Nemanja Damjanov, et al. The 2017 EULAR standardised procedures for ultrasound imaging in rheumatology. *Annals of the Rheumatic Diseases*, 76(12):1974–1979, 2017.
- [282] Jong Chul Ye and Woon Kyoung Sung. Understanding geometry of encoder-decoder CNNs. In *International Conference on Machine Learning*, pages 7064–7073. PMLR, 2019.
- [283] Saeed Miramini, David W Smith, Lihai Zhang, and Bruce S Gardiner. The spatio-temporal mechanical environment of healthy and injured human cartilage during sustained activity and its role in cartilage damage. *Journal of the mechanical behavior of biomedical materials*, 74:1–10, 2017.
- [284] David W Smith, Bruce S Gardiner, Lihai Zhang, and Alan J Grodzinsky. *Articular cartilage dynamics*. Springer, 2019.
- [285] Petri Tanska, Petro Julkunen, and Rami K Korhonen. A computational algorithm to simulate disorganization of collagen network in injured articular cartilage. *Biomechanics and modeling in mechanobiology*, 17(3):689–699, 2018.

Bibliography

- [286] François Rannou, Mathias François, Marie-Thérèse Corvol, and Francis Berenbaum. Cartilage breakdown in rheumatoid arthritis. *Joint Bone Spine*, 73(1):29–36, 2006.
- [287] Laura J Brattain, Brian A Telfer, Manish Dhyani, Joseph R Grajo, and Anthony E Samir. Machine learning for medical ultrasound: status, methods, and future opportunities. *Abdominal radiology*, 43(4):786–799, 2018.
- [288] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Communications of the ACM*, 60(6):84–90, 2017.
- [289] Veronika Cheplygina, Marleen de Bruijne, and Josien PW Pluim. Not-so-supervised: a survey of semi-supervised, multi-instance, and transfer learning in medical image analysis. *Medical image analysis*, 54:280–296, 2019.
- [290] Jiaolong Xu, Liang Xiao, and Antonio M López. Self-supervised domain adaptation for computer vision tasks. *IEEE Access*, 7:156694–156706, 2019.
- [291] David Berthelot, Rebecca Roelofs, Kihyuk Sohn, Nicholas Carlini, and Alex Kurakin. Adamatch: A unified approach to semi-supervised learning and domain adaptation. *arXiv preprint arXiv:2106.04732*, 2021.
- [292] Sicheng Zhao, Xiangyu Yue, Shanghang Zhang, Bo Li, Han Zhao, Bichen Wu, Ravi Krishna, Joseph E Gonzalez, Alberto L Sangiovanni-Vincentelli, Sanjit A Seshia, et al. A review of single-source deep unsupervised visual domain adaptation. *IEEE Transactions on Neural Networks and Learning Systems*, 2020.
- [293] Xuhong Li, Haoyi Xiong, Xingjian Li, Xuanyu Wu, Xiao Zhang, Ji Liu, Jiang Bian, and Dejing Dou. Interpretable deep learning: Interpretation, interpretability, trustworthiness, and beyond. *Knowledge and Information Systems*, pages 1–38, 2022.