UNIVERSITÀ POLITECNICA DELLE MARCHE
CORSO DI DOTTORATO DI RICERCA IN INGEGNERIA DELL'INFORMAZIONE
CURRICULUM IN INGEGNERIA INFORMATICA, GESTIONALE E DELL'AUTOMAZIONE

Towards digital patient monitoring: deep learning methods for the analysis of multimedia data from the actual clinical practice

Ph.D. Dissertation of:
Lucia Migliorelli

Advisor:
Prof. Emanuele Frontoni

Coadvisor:
Sara Moccia, PhD

Curriculum Supervisor:
Prof. Franco Chiaraluce

XX edition - new series

UNIVERSITÀ POLITECNICA DELLE MARCHE
CORSO DI DOTTORATO DI RICERCA IN INGEGNERIA DELL'INFORMAZIONE
CURRICULUM IN INGEGNERIA INFORMATICA, GESTIONALE E DELL'AUTOMAZIONE

Towards digital patient monitoring: deep learning methods for the analysis of multimedia data from the actual clinical practice

Ph.D. Dissertation of:
Lucia Migliorelli

Advisor:
Prof. Emanuele Frontoni

Coadvisor:
Sara Moccia, PhD

Curriculum Supervisor:
Prof. Franco Chiaraluce

XX edition - new series

UNIVERSITÀ POLITECNICA DELLE MARCHE
CORSO DI DOTTORATO DI RICERCA IN INGEGNERIA DELL'INFORMAZIONE
FACOLTÀ DI INGEGNERIA
Via Brecce Bianche – 60131 Ancona (AN), Italy

Abstract

Acquiring information on patients' health status from video recordings analysis is a crucial opportunity to enhance current clinical assessment and follow-up practices. This PhD thesis proposes four automatic deep learning (DL)-based systems to support the clinician in monitoring patients both in hospital and home environments.

The first proposed monitoring system is designed for preterm infants admitted to neonatal intensive care unit (NICU). Assessing preterm infants' limb-movement has a strong predictive value for early diagnosing the presence of neurodevelopmental disorders. Despite its relevance, preterms' movement monitoring in NICUs is still based on direct observation of the patient by clinicians and has, consequently, the limitation of being qualitative and discontinuous. To support clinicians, the proposed system automatically analyses depth video recordings acquired by a camera placed over the infant's crib and lays the foundation for a new paradigm of monitoring preterm infants by showing promising results with a Dice similarity coefficient (*DSC*) of 0.88 in joints detection and an inference time on a single board computing (SBC) device of 20 frames per second.

Visual assessment is carried out also in specialised centres for the treatment of children with autism. In this context, the applied behaviour analysis (ABA) operator observes the child and notes the progress made during the therapy programme. The second monitoring system supports the ABA operator in evaluating the improvements achieved in terms of behavioural autonomy by the autistic patient. The system automatically analyses images of the patient washing his/her hands using a camera installed above the sink in a specialised centre and provides clinicians with an index to quantify patient's progress and to tailor treatment protocols. Tested on data collected during clinical practice, the autonomy index predicted by the algorithm differs by one percentage point from the actual index proving to be a valuable ally for ABA operators.

The last two proposed systems support clinicians in assessing patients while rehabilitating with the smart walker and those suffering from dysarthria. The third monitoring system showed in this Thesis exploits images collected by two RGB-D cameras installed on a smart walker. Current digital assessment systems, in this clinical scenario, are limited to the characterisation of specific anatomical areas (e.g., the legs) without considering the body as a whole. The proposed DL method automatically processes the RGB-D images acquired by the cameras to derive the pose of the patient's body while using the smart walker, achieving an error of 44.05 mm. The

system is designed to be effective and return real-time results on low-cost hardware (inference time=26.6 ms) providing the clinicians with an instantaneous assessment of patient performance.

The last monitoring system involves patients suffering from dysarthria (a group of speech disorders induced by neurodegenerative diseases). The onset of dysarthria has a strong psycho-physical impact on both the patient and his/her relatives. Therefore, the immediate recognition of the functional changes that the disease causes in those who are affected with, is fundamental to allow clinicians to prescribe corrective and compensatory strategies to the communicative disability. The proposed approach aims to evaluate the impact of dysarthria in oro-facial muscles district and implements a DL algorithm that analyses RGB images of patients with amyotrophic lateral sclerosis (ALS) and stroke. The network is trained to derive the position of 68 facial landmarks and obtains a normalised mean error of 1.79, surpassing current assessment methods based on direct patient observation by neurologists and speech therapists. The value of the proposed system is twofold: it guarantees an exhaustive collection of quantitative data related to a rare pathology and it imagines a new care model for patients who, unable to reach the reference clinical centre, may use their smartphone to conduct the assessment from home.

Each approach implemented was developed and validated on multimedia data collected during the actual clinical practice. The acquisition systems have been designed to be easily deployed in home environment to allow the collection of details for enriching patient's clinical history. Each proposed system stems from the clinical need of having new tools to treat patients, able at collecting structured, easily accessible and shareable information. This research will continue to be enhanced to ensure that clinicians, who are increasingly challenged by tight working schedules, have more time to devote to patients, to treat them better and to the best of their ability.

Sommario

Acquisire informazioni sullo stato di salute dei pazienti a partire dall'analisi di video registrazioni è un'opportunità cruciale per potenziare le attuali pratiche cliniche di valutazione e follow-up. Questa Tesi di Dottorato propone quattro sistemi automatici basati su apprendimento profondo (*deep learning* (DL)) utili a supportare il clinico nel monitoraggio di pazienti sia in ambiente ospedaliero sia in ambiente domestico.

Il primo sistema di monitoraggio è pensato per i neonati prematuri ricoverati in terapia intensiva. Valutare i movimenti degli arti dei pretermine ha un forte valore predittivo per la diagnosi precoce della presenza di disturbi del neurosviluppo. Nonostante la sua rilevanza, il monitoraggio del movimento nelle terapie intensive neonatali è ancora prettamente basato sull'osservazione diretta del paziente da parte del clinico ed ha, conseguentemente, il limite di essere qualitativo e discontinuo. Per offrire supporto ai clinici, il sistema proposto analizza automaticamente le video-registrazioni di profondità acquisite da una telecamera posta sopra la culla del neonato e pone le basi per un nuovo paradigma di monitoraggio dei nati prematuri mostrando risultati promettenti con un Dice similarity coefficient (*DSC*) pari a 0.88 nel riconoscimento dei giunti e un tempo di inferenza su un device del tipo computer a scheda singola (*single board computing* (SBC)) di 20 fotogrammi al secondo.

La valutazione di tipo osservazionale non viene attuata solamente all'interno dei reparti di terapia intensiva neonatale ma anche nei centri specializzati per il trattamento dei bambini con sindrome dello spettro autistico. In questo contesto, l'operatore specializzato in analisi applicata del comportamento (ABA) osserva il bambino ed annota i progressi raggiunti nel corso del programma terapeutico. Il secondo sistema di monitoraggio supporta l'operatore ABA nella valutazione dei miglioramenti raggiunti in termini di autonomia comportamentale dal paziente autistico. Il sistema, grazie ad una telecamera installata sopra al lavandino di un centro specializzato, analizza automaticamente le immagini che riprendono il paziente nell'atto di lavarsi le mani e restituisce ai clinici un indice utile sia alla valutazione dei progressi del paziente sia alla maggior personalizzazione dei protocolli terapeutici. Testato sui dati raccolti durante la pratica clinica l'indice di autonomia predetto dall'algoritmo differisce di un punto percentuale rispetto all'indice reale dimostrandosi un alleato valido per gli operatori ABA.

Gli ultimi due sistemi di monitoraggio sono rivolti a pazienti in età adulta e sfruttano l'elaborazione automatica di immagini per supportare il clinico nel valutare i

pazienti che si riabilitano con il deambulatore smart e quelli che soffrono di disartria. Il terzo sistema di monitoraggio presentato nella tesi sfrutta le immagini raccolte da due telecamere RGB-D installate su un deambulatore *smart*. Gli attuali sistemi digitali valutativi, in questo scenario clinico, sono limitati alla caratterizzazione di aree anatomiche specifiche (ad es., le gambe) senza considerare il corpo nella sua interezza. Il metodo di DL proposto processa automaticamente le immagini RGB-D acquisite dalle telecamere e ricava la posa del corpo del paziente che cammina con il deambulatore raggiungendo un errore di 44.05 mm. Il sistema è studiato per essere efficace e restituire risultati in tempo reale su un hardware economico (tempo di inferenza= 26.6 ms) garantendo al clinico una valutazione istantanea delle performance dei pazienti.

L'ultimo sistema di monitoraggio proposto coinvolge invece pazienti affetti da disartria (insieme dei disordini dell'eloquio indotti dalle patologie neurodegenerative). L'insorgere della disartria impatta fortemente a livello psico-fisico sia sul paziente sia sui suoi congiunti per questo riconoscere immediatamente i cambiamenti funzionali che la patologia causa in chi ne è affetto è fondamentale per permettere al clinico di prescrivere strategie correttive e compensatorie alla disabilità comunicativa. L'approccio proposto nasce con l'obiettivo di valutare l'impatto della disartria nei muscoli del distretto bucco-facciale ed implementa un algoritmo di DL che analizza immagini RGB di pazienti con sclerosi laterale amiotrofica e ictus. La rete è allenata per ricavare la posizione di 68 *landmark* facciali ed ottiene un errore medio pari a 1.79 superando le attuali modalità valutative basate sull'osservazione diretta del paziente da parte di neurologi e logopedisti. Il valore del sistema proposto è duplice in quanto garantisce una esaustiva raccolta di dati quantitativi relativi ad una patologia rara e immagina un nuovo modello assistenziale per il paziente che, spesso impossibilitato a raggiungere il centro clinico di riferimento, può sfruttare il proprio smartphone per condurre la valutazione da casa.

Ogni approccio implementato è stato sviluppato e validato su dati multimediali raccolti nella pratica clinica. I sistemi di acquisizione sono stati costruiti per essere facilmente riprodotti in ambiente domestico per permettere la collezione di dettagli importanti sulla storia clinica di ogni paziente. Ogni sistema proposto nasce dall'esigenza clinica di avere a disposizione nuovi strumenti per curare i pazienti che raccolgano informazioni strutturate, facilmente accessibili e condivisibili e si svilupperà in futuro per garantire ai medici, sempre più provati dai ritmi lavorativi serrati, più tempo da dedicare ai pazienti, per curarli meglio e al meglio delle proprie capacità.

Acronyms

- Absolute $MPJPE$ (A_MPJPE)
- Accuracy (Acc)
- Adaptive moment estimation (Adam)
- Amyotrophic lateral sclerosis (ALS)
- Applied behaviour analysis (ABA)
- Artificial intelligence (AI)
- Atrous spatial pyramid pooling module (ASPP)
- Autism spectrum disorders (ASD)
- Convolutional neural network (CNN)
- Decoder (Dec)
- Deep learning (DL)
- Dense-Atrous (DeA)
- Dice similarity coefficient (DSC)
- Effective receptive field (ERF)
- Encoder (Enc)
- False positive (FP)
- Feature pyramid network extractor (FPN)
- General movement assessment (GMA)
- Height (H)
- Interquartile range (IQR)
- Mean absolute error (MAE)
- Mean Per-joint position error ($MPJPE$)
- Mean square error loss (L_{MSE})
- Neonatal intensive care unit (NICU)
- Normalized mean error (NME)

- Overlap ratio (OR)
- Per-pixel binary cross entropy loss (L_{CE})
- Percentage of correct keypoints (PCK)
- Precision ($Prec$)
- Procrustes-aligned $MPJPE$ (PA_$MPJPE$)
- Rectified linear unit (ReLU)
- Recall (Rec)
- Region of interest (RoI)
- Root mean square distance ($RMSD$)
- Single board computing (SBC)
- Stochastic gradient descent (SGD)
- System Improvement for neonatal care (SINC)
- True negative (TN)
- True positive (TP)
- Width (W)
- World health organization (WHO)

Contents

1 Background and motivation **1**
1.1 How health is changing in the digital era 1
1.1.1 From Hippocrates to deep learning: how patient care is evolving 2
1.2 Aim of the thesis 3
1.3 Thesis overview 6
1.4 Thesis contribution 7
1.5 Publications 10

2 Preterm infants' limb movement monitoring via depth-video analysis **13**
2.1 Preterm birth and the relevance of monitoring infants' limb-movement 13
2.2 The babyPose dataset 16
2.2.1 Data annotation and ethical considerations 20
2.3 Preterm infants' limb-pose estimation via spatio-temporal features analysis 21
2.3.1 Methods 23
2.3.2 Experimental protocol 27
2.3.3 Results 29
2.3.4 Discussion 36
2.4 Dense-atrous spatial-convolutional blocks to estimate preterms' limb-pose 39
2.4.1 Methods 40
2.4.2 Experimental Protocol 44
2.4.3 Results 47
2.4.4 Discussion 52
2.5 A sustainable deep learning approach for preterm infants limbs' detection 54
2.5.1 Efficiency in convolutional neural networks as a mean to improve sustainability 54
2.5.2 Methods 56
2.5.3 Experimental Protocol 58
2.5.4 Results 60
2.5.5 Discussion 62
2.6 Conclusion and future perspective 64

3 Automatic assessment of the autistic child's autonomy in daily actions 67
3.1 Background and motivation . . . 68
3.2 Methods . . . 70
3.2.1 Data acquisition protocol: the hand-washing case of study . . 70
3.2.2 Network architecture . . . 71
3.2.3 Training strategy . . . 72
3.2.4 Hand-Washing autonomy index . . . 72
3.3 Experimental Protocol . . . 72
3.3.1 Dataset . . . 72
3.3.2 Training settings . . . 73
3.3.3 Ablation study and comparison with other architectures . . . 73
3.3.4 Performance assessment . . . 73
3.3.5 Performance assessment on challenging frames . . . 74
3.4 Results . . . 75
3.5 Discussion . . . 77
3.6 Conclusion and future perspectives . . . 78

4 Estimating human pose from RGB-D images acquired via a smart walker 81
4.1 Background and motivation . . . 82
4.2 Methods . . . 83
4.2.1 Acquisition set-up . . . 83
4.2.2 Walker dataset . . . 84
4.2.3 Dataset preparation . . . 86
4.2.4 Model framework . . . 87
4.2.5 Deployment . . . 90
4.3 Experimental protocol . . . 91
4.3.1 Dataset details . . . 91
4.3.2 Implementation Details . . . 91
4.3.3 Performance metrics . . . 92
4.3.4 Model variants . . . 93
4.4 Results . . . 94
4.4.1 2D-stage . . . 94
4.4.2 Complete model . . . 96
4.4.3 Deployment . . . 97
4.4.4 Benchmark and ablation studies . . . 97
4.5 Discussion . . . 102
4.6 Conclusion and future perspectives . . . 104

5 End-to-end facial landmark detection to assess dysarthria evolution 107
5.1 Background and motivation . . . 108
5.2 Methods . . . 109
5.2.1 The Toronto NeuroFace dataset . . . 109
5.2.2 Mask-RCNN for facial landmark detection . . . 109
5.3 Experimental Protocol . . . 111
5.3.1 Training settings . . . 111
5.3.2 Ablation studies . . . 111
5.3.3 Evaluation metrics . . . 112
5.4 Results . . . 112
5.5 Discussion . . . 113
5.6 Conclusion and future perspectives . . . 114

6 Conclusive remarks 115
6.1 Conclusion . . . 115
6.2 Impact . . . 116
6.3 Future perspectives . . . 117

List of Figures

1.1 Workflow of the proposed thesis. The convolutional neural networks (CNNs)-based monitoring systems share the implementation of an acquisition set-up based on RGB-D cameras designed (i) to not hinder operators during the actual clinical practice and (ii) to not come in contact with the patient's body. 4

2.1 Depth-image acquisition setup. The depth camera is positioned at ∼40cm over the infant's crib and it does not hinder health-operator movements. 15

2.2 The workflow of the described convolutional neural networks (CNNs)-based methodologies to preterm infants' movement monitoring. The first proposed pipeline is based on 3D convolutions for depth-stream analysis. This pipeline was trained and validated on the babyPose-v1 dataset. The second designed pipeline leverages 2D convolutions to estimate limbs' pose. The third proposed approach implements a CNN to detect infants' limbs. These latter two approaches leverage the babyPose-v2 dataset. A comprehensive description of the 3 monitoring systems is provided in Sec. 2.3, Sec. 2.4, and Sec. 2.5, respectively. 17

2.3 Proof of data variability. Frames a), c), d) and e) show samples of external occlusions caused by the presence of sheets, pillows, splints, therapy equipment or the hands of the operator and parents. Frames f) and h) show samples of incorrect positioning of the acquisition set-up with respect to the crib. Frames b) and g) demonstrate variability in terms of pixel intensity level. 18

2.4 Preterm infant's joint model (including joints and joint-connections) superimposed on a sample depth frame. Inspired by clinical considerations, only limb joints are considered. LS and RS: left and right shoulder, LE and RE: left and right elbow, LW and RW: left and right wrist, LH and RH: left and right hip, LK and RK: left and right knee, LA and RA: left and right ankle. 19

2.5 Workflow of the proposed pipeline to preterm infants' pose estimation with spatio-temporal features extracted from depth videos. The input consists of a temporal clip of W_d consecutive depth frames, which are processed by two convolutional neural networks (CNNs) to roughly detect joint and joint-connection (affinity maps) and refine joint and joint-connection detection (confidence maps), respectively. 22

2.6 Ground-truth samples for the detection network. Samples are shown for (b) left knee (LK), (c) left hip (LH), and (d) their connection. . . . 23

2.7 Ground-truth samples for the regression network. Samples are shown for (b) left knee (LK), (c) left hip (LH), and (d) their connection. . . . 24

2.8 3D detection convolutional neural network (CNN) for preterm infants' joints and joint-connections position estimation. It takes in input a stack of W_d=3 depth frames and outputs 60 (W_d x 20) affinity maps. The different colours of the convolutional blocks are coded in the legend on the bottom of the figure. 26

2.9 3D regression convolutional neural network (CNN) for preterm infants' joints and joint-connections position estimation. This network takes in input the W_d=3 depth frames and the output of the detection network (i.e., 60 affinity maps) and outputs 60 (W_d x 20) confidence maps. 27

2.10 Boxplots of the recall (*Rec*) for joint (top) and joint-connection (bottom) detection achieved with the proposed 3D pipeline. Results of its akin 2D are shown for comparison, too. For colors and acronyms, refer to the joint model in Fig. 2.4. 32

2.11 Boxplots of the root mean squared distance (*RMSD*) computed for the four limbs separately. Boxplots are shown for the 2D and 3D pipeline. Asterisks highlight significant differences. 33

2.12 Sample qualitative results for pose estimation obtained with the 2D (left) and 3D (right) pipeline. White arrows highlight estimation errors, mainly due to homogeneous image intensity. 34

2.13 Sample qualitative results for challenging cases. First row: one joint was not detected due to auto-occlusion (from left to right: right shoulder, right shoulder, right hip, right hip). Second row: one or more joints were not detected due to external occlusion (from left to right: joint of the left limbs, right ankle, left arm - due to healthcare operator hand presence, and right knee and ankle - due to plaster). Last row: image noise and intensity inhomogeneities prevented joint detection. . 35

2.14 Workflow of the proposed pipeline to preterm infants' pose estimation with spatial features extracted from depth frames. The input consists of a depth frame processed by two convolutional neural networks (CNNs) (i.e., Dea-detection CNN and DeA-regression CNN) to roughly detect joint and joint-connection (affinity maps) and refine joint and joint-connection detection (confidence maps), respectively. . 41

2.15 Graphical representation of the dense-atrous-based (DeA)-detection network. The body of the network consists of 8 convolutional layers (4 layer for the encoding (Enc) path and 4 layers for the decoding (Dec) path). Each coloured block in a convolutional layer implements: (i) the convolution operation, differing for kernel sizes (3x3 or 2x2 or 1x1), (ii) the batch normalization (BN) and (iii) the activation function (Rectified Linear Unit (ReLU)). Inspired by [1], the flowing pathway between Enc2-Dec2 couples the classical long-skip connection (orange) and the DeA pathway (pink). The detail of the DeA pathway between Enc2-Dec2 (pink) is shown on the far right of the figure. The colour-coded legend is located in the lower left corner. 42

2.16 Graphical representation of the dense-atrous-based (DeA)-regression network. The input to the network consists of a stack with the input image and 20 affinity maps, which are produced from the DeA detection network, while the output consists of 20 confidence masks. The body of the network consists of 6 convolutional layers. Each violet block in a convolutional layer implements: (i) the convolution operation (kernel size= 3x3), (ii) the batch normalization (BN) and (iii) the activation function (Rectified Linear Unit (ReLU)). 43

2.17 Boxplot of the Recall (*Rec*) for joints prediction. These results were achieved by the network tested for the ablation studies summarized in Table 2.4. Blue: results achieved by the proposed DeA detection network; Red, pink and yellow: results achieved by the DeA detection-1, DeA detection-2 and the 2D detection CNN, respectively. In green the median value was reported. 48

2.18 Boxplot of the recall (*Rec*) for joints prediction. These results were achieved by the networks tested for further investigation on DeA-detection efficiency (as reported in Table 2.5). Blue: results achieved by the proposed DeA detection network; orange, red and purple: results achieved by the Asy-DeA-Enc detection, Asy-Dea-Body detection, Asy-DeA detection, respectively. In green the median value was reported. 49

2.19 Samples of qualitative results in challenging frames for the DeA detection network (second column, images marked with a blue square) and the 2D detection CNN (third column, images marked with a yellow square). The first column shows the original images while the second and the third column represent the predictions of the network (red) and the corresponding ground truth (blue) superimposed to the preprocessed images. The white arrows in the third column highlight the prediction errors committed by the 2D detection CNN while the cross stands for no predictions. 50

2.20 Boxplot of the root mean square distance error *RMSD* calculated for each of the four limbs. The violet boxplots show the results achieved by the the DeA regression network while the orange the results of the 2D regression CNN. In green the median value was reported. 51

2.21 Samples of qualitative results achieved by the DeA-regression network when estimating the limb-pose. Limb-pose was superimposed to the original depth frame. 51

2.22 Structure of detection architectures: 2D detection convolutional neural network (left), the TwinEDA (center) and the EDANet in [2]. TwinEDA combines the architectural choices of the 2D detection network (e.g., bi-branch architecture to process joints and connections in parallel) and the EDANet network (e.g., EDA modules) [2] to ensure an effective and efficient architecture. The color-caption is shown in the bottom left corner of the figure. The dashed grey arrows highlight the direction of information flow along the networks. 57

2.23 Boxplots for quantitatively evaluate the efficiency and efficacy of the network when detecting the joints. The x-axis shows the network inference rate in terms of frames per second (FPS). The y-axis shows the performance in terms of Dice similarity coefficient (*DSC*) (top) and Recall (*Rec*) (bottom) for mean-joint-detection. The different colors of the bloxplots represent the different architectures while the black line depicts the median values, the caption is shown at the bottom of the image. 59

3.1 The acquisition set-up (white square, left), placed in the bathroom to record the sink (pink square, right), consisted of an Astra Mini S-Orbbec® RGB-D camera and a minipc Intel® NUC core *i5*. 68

3.2 Workflow of the proposed deep learning (DL)-based application to classify RGB frames in which the child washes his/her hands with the help of the operator (*aid*) or autonomously (*no-aid*). 69

3.3 VGG16 neural network to classify *aid* and *no-aid* frames. In purple convolutional layers, in blue max pooling layers, in red fully connected layers. 71

3.4 Example of challenging frames: on the left side the child is aided by the operator (blue square), on the right side the child washes his hands autonomously (yellow square). 74

3.5 Confusion matrix for fine-tuned VGG16 (on the left side) and fine-tuned ResNet50 (on the right side): the two best performing models. . 75

3.6 Boxplot of the Overlap Ratio (*OR*) for the *no-aid* class in yellow and the *aid* class in blue. Median values of the two distributions are shown in red. 75

3.7 Confusion matrix of the fine-VGG16 tested on challenging frames . . 76

4.1 Summary of the method used to relate the Xsens keypoint data to the posture camera referential. The skeleton is shown in 3D along with the pointcloud from both cameras. 84

4.2 Processed input image (a) and depth (b) frames which will be fed to the model. stacked Gaussian probability keypoint (c) and connection (d) heatmaps. 85

4.3 From left to right: Outside view of the acquisition setup; collected data from concatenated RGB frames overlaid with projected 2D skeleton; Concatenated depth frames overlaid with projected 2D skeleton; merged pointcloud overlaid with 3D skeleton (data from the gait camera is transformed, through an extrinsic transformation, to the posture camera referential). 85

4.4 Proposed two-stage model framework. The 2D-stage takes the input frames and regresses the keypoints and connections heatmaps using a fully convolutional network. Soft-argmax is used on the keypoint heatmaps to obtain the 2D keypoint locations, which are lifted to 3D space using a fully connected regression network aided by the depth information. 88

4.5 2D-stage model architecture: The processed input image and depth frames are concatenated and passed by a pretrained backbone (EfficientNet-lite0), yielding multi-level features. The higher level features are enhanced by an atrous spatial pyramid pooling module and then upsampled through two branches, to produce connection and keypoint heatmaps. A refining module is used to improve keypoint heatmaps, from which the joint locations are extracted using the soft-argmax operator. 89

4.6 Chosen percentage of correct keypoints (*PCK*) threshold radius of 6 pixels. Detection values inside the circles, for each keypoint, are considered correctly detected. 92
4.7 **(a)** Keypoint heatmaps, **(b)** Connection heatmaps and **(c)** 2D keypoints and connections predicted by the 2D-stage of the model. **(d)** Corresponding 2D keypoint ground-truth labels. All data are overlaid on top of the input image frame. 95
4.8 Boxplot per-joint error (*MPJPE*) for the 2D detection, across all test frames obtained from the 2D-stage (extreme outliers were removed for better visibility). The dashed line marks the 6 pixel threshold defined. 95
4.9 Boxplot per-joint error results for the 3D detection, across all test frames, with absolute values relative to the posture camera referential, obtained with the complete model (extreme outliers were removed for better visibility). 96
4.10 Predictions obtained from the model running on the smart walker, in 2D (a) and 3D (b) spaces. Connections between the hips and legs are not rendered in 2D due to the discontinuity between camera frames. The 3D visualization used the RViz package from the ROS environment to render the 3D keypoint locations relative to the walker. The posture camera frame of reference is also displayed. 97
4.11 Results obtained when removing information from image (upper images) and depth (lower images) inputs. The image+depth inputs for each experiments are grouped in the left side, while the corresponding model predictions (keypoint heatmaps, connection heatmaps, 2D keypoints, 3D keypoints) are grouped in the right. 2D outputs were overlaid on the image input. 3D ground truth keypoints are shown overlaid with higher transparency. 98
4.12 (a) 3D skeleton obtained through keypoint depth projection using the camera Pinhole model (the leg/feet keypoints were transformed to the posture camera referential using the known extrinsic transformation). (b) The skeleton obtained after the residual correction step of the Projection_Residual method. 100

5.1 Sample image with the 68-facial landmarks (green dots) and the bounding box face annotation (black square). 109
5.2 Mask-RCNN for facial-landmark detection. 110

List of Tables

2.1 Joint-detection performance in terms of median Dice similarity coefficient (*DSC*) and recall (*Rec*). Inter-quartile range is reported in brackets. The metrics are reported separately for each joint. For joint acronyms, refer to the joint-pose model in Fig. 2.4. 30

2.2 Joint-connection detection performance in terms of median Dice similarity coefficient (*DSC*) and recall (*Rec*). Inter-quartile range is reported in brackets. The metrics are reported separately for each joint connection. For joint acronyms, refer to the joint-pose model in Fig. 2.4. 31

2.3 Limb-pose estimation performance in terms of median root mean square distance (*RMSD*), with interquartile range in brackets, computed with respect to the ground-truth pose. The *RMSD* is reported for each limb, separately. Results are reported for the 2D and 3D pipeline, as well as for the 3D detection-only, 3D regression-only and state-of-the-art architectures. 33

2.4 Ablation study for the DeA detection network. The table shows also the number of trainable parameters fot each of the architecture. 45

2.5 Comparisons for optimizing model efficiency. The table reports the names of the newly stated architectures, their trainable parameters and the architectural component where we applied asymmetric convolutions. The DeA detection network was compared against its asymmetric variations. Asy-DeA-Enc detection implements the original decoder (Dec) and DeA pathways while keeping the encoder (Enc) asymmetric. Asy-DeA-body implements the original DeA pathways while making asymmetric both the Enc and the Dec. The Asy-DeA detection represents the asymmetric version of the DeA detection. . . 46

2.6 Joint-connection- and joint- detection performance in terms of median recall (*Rec*) and Dice Similarity Coefficient (*DSC*). 47

2.7 Sustainable investigation: the performance of the TwinEDA network was compared with that of the two networks chosen as baselines for its design (i.e., EDANet [2] and the 2D detection CNN), TwinEDA-v0 and TwinEDA-v1. The table reported for each architecture the number of trainable parameters. 60

2.8 Quantitative results for joints' and connections' detection in terms of median Dice similarity coefficient (*DSC*) and Recall (*Rec*). Interquartile ranges for each measure are shown in parentheses. 61

2.9 Model efficiency assessments in terms of inference speed (i.e., the time to predict a single depth frame) for all the 5 tested architectures and memory requirements occupied by each model. 62

3.1 Dataset description: we used the annotated frames from 36 videos to train and validate the architecture (70% of frames to train and 30% of frames to validate), while the annotated frames of 1 video were used to test the architecture. 71

3.2 Results of the VGG16 and the ResNet50, both with fine-tuning technique and from scratch. Results were evaluated in terms of: class-specific classification precision ($Prec_i$), recall (Rec_i), f1-score ($f1_i$), for $i \in [aid, no\text{-}aid]$ and classification Accuracy (*Acc*). 76

4.1 Results summary of the 3D-stage and comparisons against the different variants for regression from the 2D-stage keypoints. The best results are highlighted in bold. 96

4.2 The default EfficientNet-lite0 backbone 2D-stage performance in comparison with the common ResNet models. OpenPose is also compared in terms of latency for reference (values were taken from the paper), as it is a common baseline on real-time huma pose estimation. The best results in each metric are highlighted in bold. 99

4.3 Importance of the refine module on the 2D-stage accuracy and latency. Its effect is investigated by removing the refine module and obtaining the 2D keypoint locations from the keypoint heatmaps branch, and by further removing the parallel connection heatmap branch. The best results in each metric are highlighted in bold. 99

4.4 Default 3D-stage lifting approach comparison with the 2D keypoint locations projection using the pixel depth and the camera Pinhole model (Projection_Raw), and by further processing using the residual correction neural network in the Projection_Residual method. The ground truth 2D keypoint locations were used as input to the regression in all methods. The best results in each metric are highlighted in bold. 100

4.5 Result summary of the default 3D-stage and lifting variants when receiving as input the ground truth 2D keypoint locations. The default 3D-stage error when regressing from the 2D-stage predictions is shown as reference. The best results in each metric are highlighted in bold. 101

4.6 2D-stage results obtained using the temporal model with 4 sequential frames, compared to the default single-frame version. 101
4.7 3D results obtained using the temporal model with 4 sequential frames, compared to the proposed single-frame version, from the 2D keypoint location predictions of the 2D-stage and ground truth. The best results in each metric are highlighted in bold. 102

5.1 Conducted ablation studies for validating the proposed architecture for facial landmark detection in patients suffering from neurodegenerative diseases. 112
5.2 Results in terms of Normalized Mean Error (*NME*) for each of the 4 tested convolutional neural network (CNNs). 112

Chapter 1

Background and motivation

1.1 How health is changing in the digital era

We are living through years of continuous change driven by technology and continuous re-invention and re-interpretation of what surrounds us in an increasingly modern and digital key. This technology is "the accelerator of humanity" [3] and, as consequence, has triggered a real revolution. This revolution is inexorable and its impact in certain scenarios depends on the attitude with whom we embrace it: we can oppose or we can adapt and ride the wave of change. This is even more true if we talk about artificial intelligence (AI) applications in the field of medicine and human health, a domain where innovative technologies are becoming pervasive. Every day new applications are devised and the technologies we dispose are continuously refined or surpassed. As a consequence, new visions and interpretations of how AI can transform the concept of the healthcare ecosystem arise. Some of these visions may be fright and hinder this revolution. These are the visions that perpetuate the dominance of the machine in the human-machine relationship. The visions that see algorithms replacing health professionals and patients interacting with detached, dehumanised applications [4].

There is, however, another vision, the one that is prepared to accept the "complex and amazing convergence between man and machine" [5] and that interprets technologies for what they actually represent: tools, designed and programmed by man to support him in his daily life. In this vision, AI and in particular deep learning (DL), that learns from data, permeate healthcare with knowledge, make it more human, and enable it to regain a level of empathy never reached in the last century [6]. Indeed, DL-based systems that integrate and improve human performance, of all, have the benefit of extending the clinician's operational capacity: supporting him/her in performing repetitive tasks and, consequently, allowing him/her to devote more time to patients, to deal with patients, to be able to treat more patients and do it better [7].

In addition to the undeniable advantages for the healthcare system, mainly related to the optimisation of existing resources (and, consequently, the reduction of personnel costs), we have to count the benefits for patients [8]. DL in health care contributes decisively to the transformation of medicine into precision-medicine, which is emerging as the medicine of the future. A medicine capable of promptly intercepting the health

needs of the patient, thanks to the ability of algorithms to process large amounts of data through integrative tools, and to support the clinician in prescribing therapies and planning a follow-up tailored to the needs of the individual, who becomes an active subject and actively involved in his/her care plan [9].

1.1.1 From Hippocrates to deep learning: how patient care is evolving

Disease is no longer the focus of Hippocrates' thinking and treatment, but men in their entirety and wholeness. In order to make an accurate diagnosis, prescribe the correct treatment and take the best possible care of the patient, Hippocrates observed, studied and considered all aspects of human life: from food to symptomatology via the analysis of the patient's living context and social status. This observation-based medicine has long been the epistemological guiding principle of the medical profession [10].

As medicine has progressed, the Hippocratic method has evolved into evidence-based medicine: a patient-centred medicine that saw in the clinician-patient dialogue its greatest potential. Evidence-based medicine follows a sequential operational methodology: (i) it tracks down clinical research and/or its systematic reviews as efficiently as possible as to have the best possible "evidence" capable of answering the formulated questions, (ii) it critically evaluates the intrinsic validity of the available clinical research and the applicability of the results to the patient under examination (iii) it orients its decisions taking into account the needs of the patient and the available evidence (iv) it makes explicit the decision and the scientific motivations that justify it [11]. The approach of contemporary evidence-based medicine is slowly changing with the introduction and use of new exponential technologies such as DL. This new medicine is still evidence-based, but it is no longer based on what is evident to the clinicians, but on the evidence that DL algorithms capture from large quantities of data [10].

The realm of health has seen an explosive growth in the volume of multi-source patients-related-information. Advances in technology have created -and will continue to create- an increasing ability to collect relevant health-related measurements from the individual patients everywhere with consumer devices. This lead to thousands or even millions of unstructured and hybrid measurements collected on a daily basis [12]. This demographic, clinical and person's lifestyle data, if interpreted correctly, as Hippocrates pointed out, contributes to a better understanding of patient's health status or even clinical condition [13].

In this scenario, DL may become a valuable ally of clinicians whose analytical capabilities are not sufficient to cope with the masses of data from patients. With respect to the more conventional methods of data analysis (e.g., standard statistical methods or machine learning methods based on handcrafted features), DL algorithms are able to detect, classify and quantify even the most imperceptible of recurring patterns within

the unstructured data they analyse [14] [15]. These algorithms empowers clinicians' ability and offers them the opportunity to perceive the advent of signs long before disease appear and to prevent the evolution of a patients' condition before a possible aggravation. This would allow clinicians to timely intervene as to prescribe corrective and compensatory therapies, leading to considerable benefits to patients and all the actors involved in the individual's care and assistance pathway [16].

These new technologies offer new possibilities that require to completely rethink the way people relate to health. Following this paradigm change, research has the role of becoming the engine of innovation. Researchers in the field of DL should imagine and promote advanced strategies for making the actual healthcare model in line with the digital age we are living in. They must propose increasingly innovative solutions to enable sustainable progress in healthcare, empower the clinician-patient relationship making it more active and bidirectional and improve all individuals quality of life breaking down eventual economic and social barriers [17].

Our increasingly digital society tends to tell its story in images, even the few remaining words, in the near future, will give way to images as a mean for messages and suggestions. Just browsing the web and social networks allows to understand the relevance of images both as a way to communicate and as a source of knowledge about a user or samples of users [18]. All data generated on the browser when analysed provide knowledge. From a photograph posted online, algorithms are able to re-identify the portrayed people [19], assess their sentiment [20], their eating habits and their lifestyle [21]. This analysis would allow to assess user behaviours and passions so as to personalise their browsing experience.

This concept of personalisation, which is already well-established in many fields of research, has yet to become a pillar in a medicine which is still too much based on direct observation of the patient by trained clinicians [22]. In this respect, DL has already proved to be effective in enhancing diagnostics through automated image interpretation, for example, in dermatology, radiology, ophthalmology and pathology [15]. However, too little research efforts have been spent on designing innovative patient-monitoring solutions to allow new clinically-relevant parameters to be collected and integrated with pre-existing knowledge about a patient.

1.2 Aim of the thesis

Considering the possibility of acquiring information about patients' health status, from the analysis of snapshots and video-recordings a crucial opportunity for the renewal of medicine particularly in the context of patient's monitoring and follow-up, the thesis proposed four **automatic and non-invasive DL-based patient-monitoring systems**. These systems, designed and tested during the actual clinical practice, exploit the potential of DL to analyse RGB and RGB-D video recordings. Each system has been designed to not hindering healthcare operators during their practice or not being in

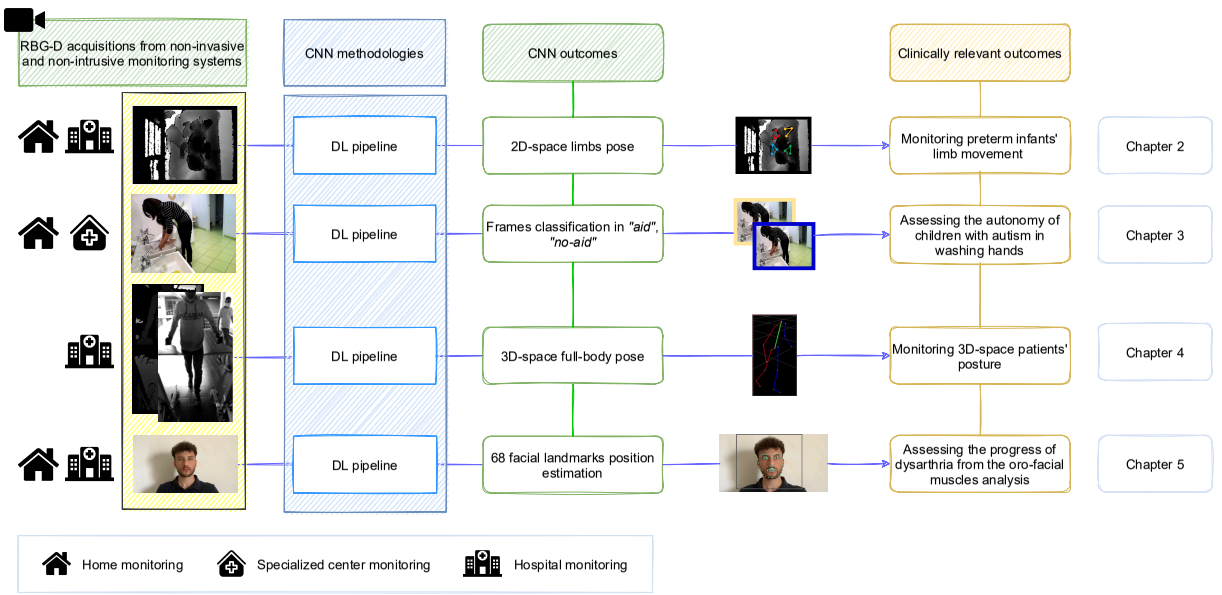


Figure 1.1: Workflow of the proposed thesis. The convolutional neural networks (CNNs)-based monitoring systems share the implementation of an acquisition set-up based on RGB-D cameras designed (i) to not hinder operators during the actual clinical practice and (ii) to not come in contact with the patient's body.

direct contact with a person's body.

The implemented DL-based applications that will be presented in this work are designed to monitor users of different ages, from infants born prematurely to adults with neurological disorders and adolescents with autism spectrum syndrome (ASD). They all stem from three main clinical needs: (i) making patients' monitoring quantitative and continuous, (ii) imagining more open and dialogue systems to quantitative data-collection and (iii) supplementing knowledge about a patient with innovative data that the patient can possibly collect on his/her own. The workflow of the proposed methodologies is showed in Figure 1.1.

The first proposed DL-based monitoring system results from a three-year research conducted in the neonatal intensive care unit (NICU) of the "G. Salesi" Hospital in Ancona (Italy). This study arises from the need to propose an innovative approach to monitor the movement of the limbs of an infant born prematurely as a way to timely detect the presence of neuro-developmental disorders [23]. The research develops over the course of the 3 years proposing increasingly effective and efficient DL models and culminates in the development of among the first single-board computing (SBC) monitoring system thought for guaranteeing accuracy in prediction and sustainability both in terms of energy consumption and costs. The value of this research is two folds: (i) it provides advanced and non intrusive monitoring support in scenarios where inpatient assessments mainly relies on direct observations and (ii) it echoes the need to develop increasingly efficient models deployable in cost-effective devices, as a key to make these advanced monitoring solutions globally accessible.

Moving on through the ages, personal autonomy skills are among the elements that mostly affect the quality of life of the child with ASD. Hand-washing, of all the basic autonomies, is crucial for the safety of the children and to improve their social integration. The applied behavior analysis (ABA), is a validated therapy for the treatment of the ASD-child's and, above all, has the benefits of improving children's social, communication, and learning skills [24]. During the application of the ABA program, operators evaluate the child's progress through observational methods, with the drawbacks of being: subjective and non-repeatable. To overcome the limitations posed by perspective evaluations, this research proposes a DL methodology that analyses RGB images from a camera mounted over the bathroom sink of the "Orizzonte" centre (Macerata, Italy), specialized in ABA therapy. The algorithm automatically classifies whether the ASD child washes his/her hands autonomously or with the support of the ABA operator and produces and index of autonomy whose evolution may be relevant to monitor the progress made by each child. This research lays the foundation for the development of a broader framework as to offer all the possible support to the ABA operators in proposing increasingly patient-specific programmes.

From teenagers to adults, the research conducted in cooperation with the University of Minho (Braga, Portugal) was aimed at improving the current monitoring methodology of adults who undergo a rehabilitation program using smart walkers. This research

proposes a DL approach to analyse multimedia data from two cameras mounted over the smart walker as to estimate the pose of the patient's whole-body in the 3D space. The dual relevance of this research lies in: (i) the development of among the first 3D-pose estimation system for the smart walker, providing clinicians with quantitative measures to establish patient-specific rehabilitation programs (ii) the proposal of an effective and efficient system, proving, once again, that computational efficiency and effectiveness are key aspects that can and should be pursued as one.

The latest research proposal targets adults suffering from neurodegenerative diseases (such as amyotrophic lateral sclerosis (ALS) or stroke). One of the most devastating effects of ALS and stroke is dysarthria, which is the set of speech disorders mainly induced by these neurodegenerative diseases [25]. Automatic evaluation of dysarthria evolution through non-invasive oro-facial muscles assessment is relevant to allow clinicians of readily identifying the instant for prescribing compensation strategies to communicative disabilities. This research proposes a DL methodology to estimate the position of 68 facial-landmarks from RGB images acquired during the outpatients assessments. From the location of these landmarks, dysarthria-evaluation indexes can be estimated as to provide clinicians with quantitative and continuously obtainable measures.

1.3 Thesis overview

An overview of the thesis structure is proposed hereafter for the sake of readability:

- **Chapter 2**: presents the crucial clinical need of making continuous and quantitative preterm infants' movement monitoring in NICUs as a way of timely recognising the presence of neuro-motor disorders. Within chapter subsections, the challenges behind the monitoring task will be analysed and a number of DL methodologies will be proposed to gradually meet the actual needs. All the presented methods have a common root: they are all based on a pipeline consisting of two convolutional neural networks (CNNs) the first, namely detection CNN, is aimed at roughly estimating the position of limb-joints and connections in space while the second, namely regression CNN, refines the previously found predictions. The proposed approaches were implemented to process depth clips or frames as to preserve ward's privacy. This chapter will further give details on the babyPose dataset-v1 and -v2 (on which the algorithms were trained and tested): the largest publicly available annotated dataset of preterm infants' depth frames acquired in the actual clinical practice [26].
- **Chapter 3** highlights the need to implement computer-based support systems for ABA operators who currently assess the progress of children with ASD through direct observation of the child coupled with paper-and-pencil assessment scales. The chapter will present the use case of the autonomy of hand-

washing (since among all the basic autonomies is crucial to preserve healthy habits) proposing a system based on DL to quantify from RGB images analysis if the child washes his/her hands alone or with the support of the ABA operators.

- **Chapter 4** aims to exploit an efficient DL-based pipeline strategy to estimate the whole-body human pose in the 3D space. The pipeline couples a first CNN to estimate the 2D-space human pose and a second network to regress the joints' coordinate in the 3D space. The DL pipeline analyses RGB-D frames resulting from the combination of frames from two distinct cameras mounted on a smart walker. This monitoring system is relevant as to promptly offer clinicians a quantitative assessment of a person's posture as to implement a preventive rehabilitation strategies to eventual impairments.

- **Chapter 5** aims to present an innovative quantitative assessments systems for dysarthria evolution evaluation. Indeed, despite its relevance, dysarthria monitoring is still based on the patients' observation by trained clinicians combined with the compilation of rating scales. This qualitative assessment method does not allow to perceive subtle changes in patients performance and consequently slows down the prescription of compensatory strategies to communicative disability. To solve the need it will be proposed an end-to-end DL pipeline to estimate the position of 68-facial landmarks from RGB frames analysis, as a way to monitor the impact caused by dysarthria progress on oro-facial muscles.

- **Chapter 6** offers an overview of the conclusions of each work from previous chapters. Then, final considerations and open challenges of healthcare ecosystem are discussed.

The chapters (i.e., chapter. 2 ÷ chapter. 5), which differs for the clinical need to be solved i) give the reader and overview of the state of the art in the field; ii) present the adopted dataset; iii) justify the choice of the proposed DL pipelines; iv) present the experimental setup and evaluation metrics ; v) provide the results for evaluating the performance of the proposed method; vi) discuss the obtained results, highlight the limitations and conclude with the future perspective of the research.

1.4 Thesis contribution

When dealing with preterm infants, monitoring limb movement is crucial to assess the presence of neuro-motor dysfunctions. During the three years of the doctorate, the following publications contributed to expanding the state of the art in the field of fully-automatic clinical support system for non-intrusive preterm infants' movement monitoring from depth video-recordings. The contribution, in journals and conferences, focused on (i) gradually improving the DL-methodologies for depth frames and clips

processing both in terms of efficacy and efficiency (ii) extensively describing the babyPose dataset for enabling other researchers in the field to reproduce the experiments and participated in the challenge of always proposing innovative support systems for clinicians in NICUs, (iii) designing the cloud-based platform that ensures the clinician can view the results of the analysis from the algorithms. The proposed research falls within the framework of the System improvement for neonatal care (SINC) project[1] which involves collaboration between universities, companies and the hospital as to propose the first fully technological crib for preterm infants.

- S. Moccia, **L. Migliorelli**, V. Carnielli and E. Frontoni, ≪Preterm Infants' Pose Estimation With Spatio-Temporal Features≫. In IEEE Transactions on Biomedical Engineering (2020).
- **L. Migliorelli**, S. Moccia, R. Pietrini, V. Carnielli and E. Frontoni. ≪The babyPose dataset≫. In: Data in brief 33 (2020), p. 106329.
- **L. Migliorelli**, E. Frontoni, S. Moccia ≪An accurate estimation of preterm infants' limb pose from depth images using deep neural networks with densely connected atrous spatial convolutions≫. In: Expert System With Applications (2021).
- **L. Migliorelli**, A. Cacciatore, V. Ottaviani, D. Berardini, R. L. Dellaca', E. Frontoni, S. Moccia ≪TwinEDA: a sustainable deep-learning approach for limb-joint detection in preterm infants' depth images≫. In: IEEE Journal of Biomedical and Health Informatics (2021).
- S. Moccia, **L. Migliorelli**, R. Pietrini and E. Frontoni, ≪Preterm infants' limb-pose estimation from depth images using convolutional neural networks≫. In: IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB) (2019).
- **L. Migliorelli**, A. Cenci, M. Bernardini, L. Romeo, S. Moccia and P. Zingaretti. ≪A Cloud-Based Healthcare Infrastructure for Neonatal Intensive Care Units≫. In: Proceedings of the ASME 2019 International Design Engineering Technical Conferences and Computers and Information in Engineering Conference. Volume 9: 15th IEEE/ASME International Conference on Mechatronic and Embedded Systems and Applications. Anaheim, California, USA (2019).
- **L. Migliorelli**, S. Moccia, G. P. Cannata, A. Galli, I. Ercoli, L. Mandolini, V. Carnielli and E. Frontoni. ≪A 3D CNN for preterm-infants' movement detection in NICUs from depth streams≫. In: Seventh National Congress of Bioengineering. Gruppo Nazionale di Bioingegneria (2021).

[1] https://www.regione.marche.it/Entra-in-Regione/Fondi-Europei/FESR/Programma-Operativo-Por-FESR

- **L. Migliorelli**, E. Frontoni, S. Appugliese, G. P. Cannata, V. Carnielli and S. Moccia. ≪Improving Preterm Infants' Joint Detection in Depth Images Via Dense Convolutional Neural Networks≫. In: 43rd Annual International Conference of the IEEE Engineering in Medicine and Biology Society (2021).

- **L. Migliorelli**, D. Berardini, F. Rossini, E. Frontoni, V. Carnielli and S. Moccia. ≪Asymmetric Three-dimensional Convolutions For Preterm Infants' Pose Estimation≫. In: 43rd Annual International Conference of the IEEE Engineering in Medicine and Biology Society (2021).

- M. Carbonari, G. Vallasciani, **L. Migliorelli**, E. Frontoni e S. Moccia. ≪End-to-end semantic joint detection and limb-pose estimation from depth images of preterm infants in NICUs≫. In: 2021 IEEE Symposium on Computers and Communications (ISCC) (2021).

Assessing the autonomy of children with ASD during the tasks of everyday life (e.g., washing hands, brushing teeth) is crucial as to enable the ABA operators to design children-specific programmes while capturing data relevant to improve and expand knowledge about this syndrome. The following publications participated in the proposal of the first non-intrusive and fully automatic autonomy assessment system for children with ASD. The system relies upon the same RGB-D camera mounted over the preterm infants' crib. In this scenario the camera was installed over the bathroom sink to film children with ASD washing their hands. Frames from RGB videos were manually annotated and used to design a DL methodology suitable for quantifying the level of autonomy of children in washing their hands. The research was conducted within the COMEACASA[2] project which brings together multidisciplinary expertise, from university to industries through social cooperatives, to design innovative care models for supporting children with ASD and their families.

- D. Berardini, **L. Migliorelli**, S. Moccia, M. Naldini, G. D. Angelis and E. Frontoni, ≪Evaluating the autonomy of children with autism spectrum disorder in washing hands: a deep-learning approach≫ In: IEEE Symposium on Computers and Communications (ISCC) (2020), pp. 1-7.

- D. Berardini, **L. Migliorelli**, S. Moccia, and E. Frontoni ≪On-the-edge systems for home monitoring of the progress of children with autism spectrum syndrome≫ In: Expert System with Applications [in submission].

Care and assistance of the elderly become essential to ensure the well-being of an increasingly elderly population. Appropriate rehabilitation strategies aimed at preserving the wellness of this population for as long as possible are crucial to meet the challenges of ageing. With the view to innovate rehabilitation strategies while offering

[2] https://www.ilfarosociale.it/tag/come-a-casa/

support to clinicians in the field, the following contribution deals with the proposal of among the first smart walker non-intrusive monitoring system. The system proposes a real-time full-body pose in the three-dimensional space. The pose is estimated from the RGB-images collected by two cameras mounted above the smart walker. The system, resulting from the dialogue between universities and hospital, provides for among the first time in literature, full-body quantitative data while caring for limiting the required computational resources.

- M. Palermo, S. Moccia, **L. Migliorelli**, E. Frontoni and C.P. Santos. ≪Real-time human pose estimation on a smart walker using convolutional neural networks≫. In: Expert Systems with Applications (2021).

The latest contributions are again the result of a collaborative dialogue between university and hospital and resulted in the establishment of the start-up and spin-off AIDAPT. They are developed within the broader "Homely Care" [3] project which was born to imagine and develop the first dysarthria remote-monitoring system for patients suffering of neurodegenerative diseases. The system analyses audio and video recordings, that the patient independently acquires through a web application, via DL algorithms to identify early signs of impairment.

- **L. Migliorelli**, F. Alborino, M. Coccia, L. Villani, E. Frontoni, and S. Moccia. ≪End-to-end facial landmark detection to characterise oro-facial impairments in neurological patients: towards innovative techniques for the assessment of dysarthria≫. In: the 17th International Conference on Computational Intelligence Methods for Bioinformatics and Biostatistics (2021).

- L. Scoppolini Massini, **L. Migliorelli**, E. Frontoni, and S. Squartini ≪A deep learning-based method for counting phoneme repetitions during a diadochokinesis task for the ALS patient≫. In: IEEE Transactions on Biomedical Engineering [in submission].

- **L. Migliorelli**, F. Alborino, E. Frontoni, S. Moccia, L. Villani, and M. Coccia. ≪Clinical validation of the Homely Care system in ASL patients≫. In: Neurological sciences [in submission].

1.5 Publications

The following publications, which are only partially related to the topic of the doctorate and will not be discussed in the thesis, result from intra- and inter-VRAI research group collaborations:

[3] https://aidaptsrl.com/en/products/homelycare/

- C. Calamanti, S. Moccia, **L. Migliorelli**, M. Paolanti, and E. Frontoni. ≪Learning-Based Screening of Endothelial Dysfunction From Photoplethysmographic Signals≫ In: Electronics (2019).

- D. Berardini, S.Moccia, **L. Migliorelli**, I. Pacifici, P. Di Massimo, M. Paolanti, and E. Frontoni, ≪Fall detection for elderly-people monitoring using learned features and recurrent neural networks≫. Experimental Results (2021).

- L. Antognoli, S. Moccia, **L. Migliorelli**, S. Casaccia, L. Scalise, and E. Frontoni. ≪Heartbeat Detection by Laser Doppler Vibrometry and Machine Learning≫ In: Sensors (2020).

- E. Frontoni, L. Romeo, M. Berardini, S. Moccia,**L. Migliorelli**, M. Paolanti, A. Ferri, P. Misericordia, A. Mancini, and P. Zingaretti. ≪A Decision Support System for Diabetes Chronic Care Models Based on General Practitioner Engagement and EHR Data Sharing≫. In: IEEE Journal of Translational Engineering in Health and Medicine (2020).

- M. Salati, **L. Migliorelli**, S. Moccia et al. ≪A Machine Learning Approach for Postoperative Outcome Prediction: Surgical Data Science Application in a Thoracic Surgery Setting≫. World Journal of Surgery 45, 1585–1594 (2021).

- S. Casaccia, R. Naccarelli, S. Moccia, **L. Migliorelli**, E. Frontoni, and G.M. Revel. ≪Development of a measurement setup to detect the level of physical activity and social distancing of ageing people in a social garden during COVID-19 pandemic≫ In: Measurement (2021).

- **L. Migliorelli**, S. Moccia, I. Avellino, M. C. Fiorentino and E. Frontoni, ≪MyDi application: Towards automatic activity annotation of young patients with Type 1 diabetes≫, 2019 IEEE 23rd International Symposium on Consumer Technologies (ISCT) (2019).

- M. Bernardini, A. Ferri, **L. Migliorelli**, S. Moccia, L. Romeo, S. Silvestri, L. Tiano, and A. Mancini ≪Augmented Microscopy for DNA Damage Quantification: A Machine Learning Tool for Environmental, Medical and Health Sciences≫. Proceedings of the ASME 2019 International Design Engineering Technical Conferences and Computers and Information in Engineering Conference. Volume 9: 15th IEEE/ASME International Conference on Mechatronic and Embedded Systems and Applications. Anaheim, California, USA. August 18–21, (2019).

- S. Moccia, L. Romeo, **L. Migliorelli**, E. Frontoni, and P. Zingaretti ≪Supervised CNN Strategies for Optical Image Segmentation and Classification in Interventional Medicine≫. In: Nanni L., Brahnam S., Brattin R., Ghidoni S., Jain L.

(eds) Deep Learners and Deep Learner Descriptors for Medical Applications. Intelligent Systems Reference Library (2020).

Chapter 2

Preterm infants' limb movement monitoring via depth-video analysis

Infants' monitoring in crib has a 100-year-old tradition. The babymonitor entered our homes many years ago as the "son" of the old walkie-talkie radios. The tool does not replace the presence of the parent, but is a valid support for the home monitoring of the newborn. The babymonitor is a constantly evolving technology that follows the changing culture and social habits. In today's families managing a baby may be complex. The need for extra support in managing such a hectic life stimulates mankind to gradually create tools that are increasingly able to respond to the pressing need. For this reason, the babymonitor has evolved from transmitting audio only to showing video of the baby.

Monitoring preterm infants' in NICU, and in particular their limbs movement, is of crucial clinical importance. A babymonitor that transmits video alone to clinicians is not enough as it needs someone to constantly review the videos and take notes on the possible events to look out for. To overcome limitations posed by sporadic and qualitative observations, the following chapter tells the story of a three-year research conducted in the NICU department of the "G. Salesi" in Ancona (Italy).

The research is inspired by the babymonitor story and imagines its technologically advanced version which embeds DL-methodologies to automatically monitor preterm infants' limbs movement with the view to support NICU healthcare team.

2.1 Preterm birth and the relevance of monitoring infants' limb-movement

Preterm birth is defined by the World Health Organization (WHO)[1] as a birth before thirty-seven completed weeks of gestation [27]. Across 184 countries surveyed, the

[1] `https://www.who.int/news-room/fact-sheets/detail/preterm-birth`

rate of preterm birth ranges from 5% to 18% of births. From 7% to 9% of pregnancies are not completed by the 40^{th} gestational week. 1% of the infants born very preterm [28], thus before the 32 week of gestation while, the 0.5% born even before the 28^{th} week (and these infants are recognised as extremely preterm infants [28]). In almost all high-income Countries, complications of preterm birth are the largest direct cause of neonatal deaths, accounting for the 35% of the world deaths a year [29].

The infants who survive have to face a wide range of morbidities associated with prematurity, with the frequency and severity of adverse outcomes rising with decreasing gestational age and quality of care [27]. Indeed, preterm birth implies immaturity of many organs and apparatuses, resulting in difficulty to cope with the extra-uterine environment [29]. Compared to infants born on time, preterm infants may face serious long- and short-term health issues mainly affecting brain, lungs and vision [30].

Timely identifying the infants at risk of developing neurobehavioural disorders is still a challenge. The Prechtl general movement assessment (GMA) has been recognized as a valuable diagnostic and prognostic tool for recognising infants at risk for neuromotor deficits [31]. However, despite its relevance, GMs follow-up still mostly relies on qualitative and sporadic observation of infants' limbs directly at the crib-side. Beside being time-consuming, this monitoring procedure (i) is discontinuous, (ii) may be prone to inaccuracies due to clinicians' fatigue and susceptible to intra- and inter-clinician variability and (iii) is limited to the time when the infant is hospitalised in NICU [32]. Sometimes this direct observation is coupled with qualitative paper-and-pencil rating scales which, additionally, are unable to perceive subtle changes in infants' performance [33, 34]. This further results in a lack of documented quantitative parameters to improve infants' care plans [35].

To solve the issue of such a perspective evaluation, in the past decades, a number of computer-based approaches was developed to support clinicians in monitoring infants' limb. In [36] and [37], wearable sensors placed on wrists and knees are used, respectively. Data from tri-axial accelerometer, gyroscope, and magnetometer (integrated in the sensor) are processed to monitor infants' limb movement via a threshold-sensitive filtering approach, achieving encouraging results. Similarly in [38], jumpsuits with embedded wearable sensors are proposed for infants' posture and movement monitoring. However, practical issues may arise when using wearable sensors. Hence, even though miniaturized, these sensors are directly in contact with the infants, possibly causing discomfort, pain and skin damage while hindering infant's spontaneous movements [39].

In the last couple of years, camera sensors have become valuable allies to overcome the aforementioned limitations. Preliminary results are achieved in [40], [41] and [42] for infant's whole-body segmentation with threshold-based or semi-supervised algorithms. However, as highlighted in [43], monitoring each limb individually is crucial to assist clinicians in the health-assessment process and semi-supervised approaches are hard to be brought to the actual monitoring practice.

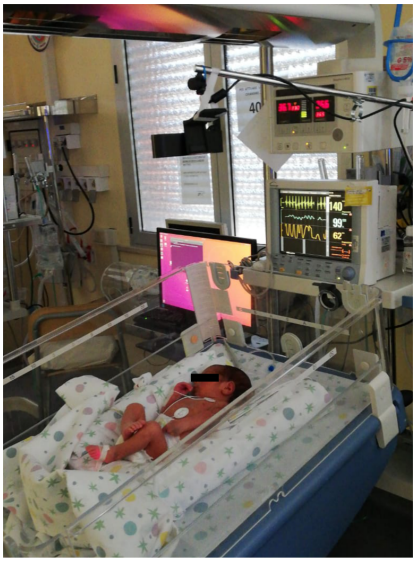


Figure 2.1: Depth-image acquisition setup. The depth camera is positioned at ∼40cm over the infant's crib and it does not hinder health-operator movements.

The issue has been tackled in [44, 45, 46]. Authors in [45] proposed a motion tracking algorithm for infants' limb-movement monitoring while [44, 46] implement a CNN for preterm infants' pose estimation via RGB frames. Despite their automatic nature all these studies leverage RGB images which may pose concerns relevant to the infants', operators' and parents' privacy.

Gathering the clinical need to make continuous and quantitative preterm infants' movement monitoring while preserving infants' and ward privacy, the following sections present a research aimed at proposing different approaches based on CNNs to estimate preterm infants' limb-pose from depth clips or single-frames. Indeed, as claimed in [47], assessing preterm infants' limb-pose is a relevant research problem with a view to automatize preterm infants' GMs monitoring. The proposed approaches were validated on data collected via a camera placed over the cribs in the NICU of the "G.Salesi Hospital", (Ancona, Italy). The data were acquired during the actual clinical practice and manually annotated with the support of trained clinicians [26]. Figure 2.1 shows the acquisition set-up which was designed to not hinder healthcare operators and parents while interacting with the infants. The camera is placed approximately 40 centimetres above the infant.

The final pipelines result from: (i) the necessity to gradually satisfy clinical (i.e., continuously quantifying preterm infants' single-limb movement) and technical needs (i.e., developing an effective and efficient monitoring system translatable into clinical practice) (ii) extensive ablation studies and comparisons against other state-of-the-art CNNs in closer fields (e.g., [48], [49]).

Further details on the following sections are provided hereafter and showed in Fig. 2.2:

- **Sec. 2.2**: extensively describes the babyPose dataset used to develop the CNN models. The dataset has two versions, the first comprising 16000 annotated depth frames while the second with 27000 frames.
- **Sec. 2.3**: shows the first CNN-based pipeline implemented which follows the clinical need of developing a system to automatically and accurately monitor single-limb movement from depth-clips analysis. The pipeline leverages two consecutive CNNs (i.e., a main 3D detection sub-network to roughly detect joints and joint-connections position and a 3D regression sub-network to refine previously found predictions) and a joint-linking step to trace limb skeleton. The CNNs implement 3D convolutions for the analysis of spatio-temporal features.
- **Sec. 2.4**: describes a revised version of the previously proposed pipeline. The pipeline consists on dense-atrous (DeA) detection CNN, a DeA regression CNN and the last joint-linking step. Both the CNNs implement 2D convolutions to analyse spatial features only. The need to propose this new methodology based on 2D convolutions-only, stems from the prohibitive computational complexity of the 3D convolutions, that were found to be unsuitable for being ported in the actual clinical practice.
- **Sec. 2.5**: presents a new efficient and effective CNN (namely TwinEDA) to detect infants' joints and joint-connections. TwinEDA analyses depth frames and results suitable for being deployed on SBC device. This research has a twofold value: (i) it demonstrates that effectiveness and efficiency can and should be pursued as one (ii) it raises the need to exploit cost-effective devices as to distribute these advanced monitoring systems in scenarios where highly-demand computational resources are not available.

Sections 2.3, 2.4, 2.5 are organized as follows: a paragraph listing the innovative content of each approach and the rationale for implementing it introduces each section. Then a method subsection follows the introduction to present each implemented architecture in detail. The experimental protocol then provides the training details, ablation studies and comparison with other architectures as to enable other researchers in the field to reproduce the experiments. This subsection is followed by the results subsection and by a broader discussion of these.

2.2 The babyPose dataset

The babyPose dataset was collected, annotated and used to implement and validate the first algorithms in the literature for infant limb-pose estimation in NICUs. This

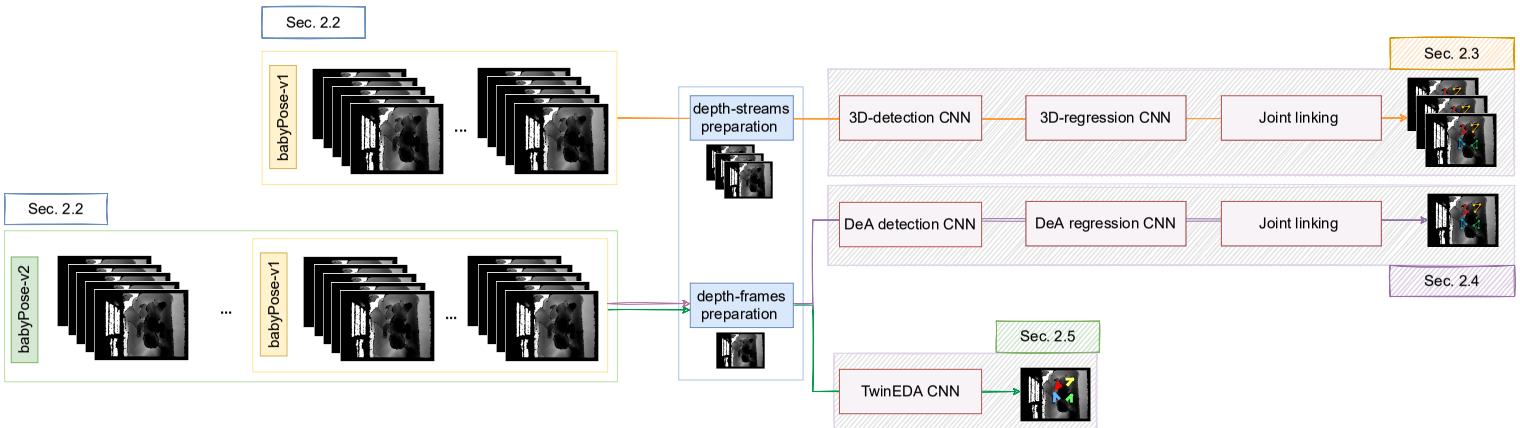


Figure 2.2: The workflow of the described convolutional neural networks (CNNs)-based methodologies to preterm infants' movement monitoring. The first proposed pipeline is based on 3D convolutions for depth-stream analysis. This pipeline was trained and validated on the babyPose-v1 dataset. The second designed pipeline leverages 2D convolutions to estimate limbs' pose. The third proposed approach implements a CNN to detect infants' limbs. These latter two approaches leverage the babyPose-v2 dataset. A comprehensive description of the 3 monitoring systems is provided in Sec. 2.3, Sec. 2.4, and Sec. 2.5, respectively.

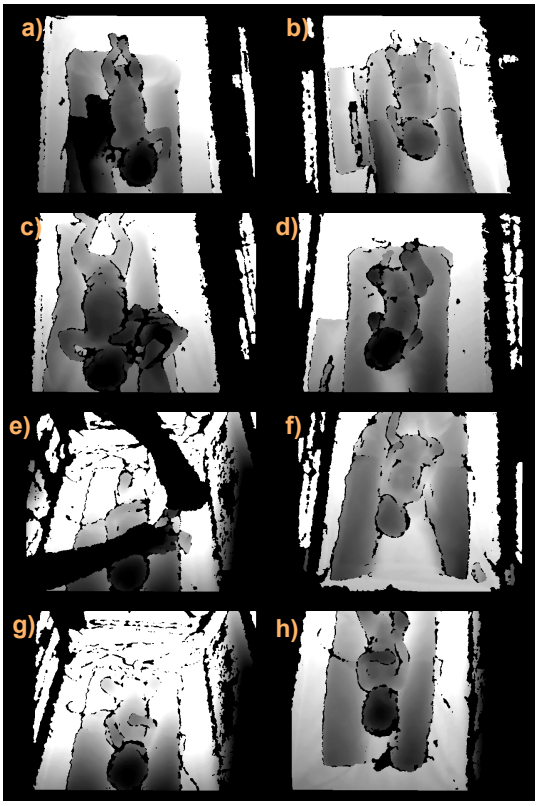


Figure 2.3: Proof of data variability. Frames a), c), d) and e) show samples of external occlusions caused by the presence of sheets, pillows, splints, therapy equipment or the hands of the operator and parents. Frames f) and h) show samples of incorrect positioning of the acquisition set-up with respect to the crib. Frames b) and g) demonstrate variability in terms of pixel intensity level.

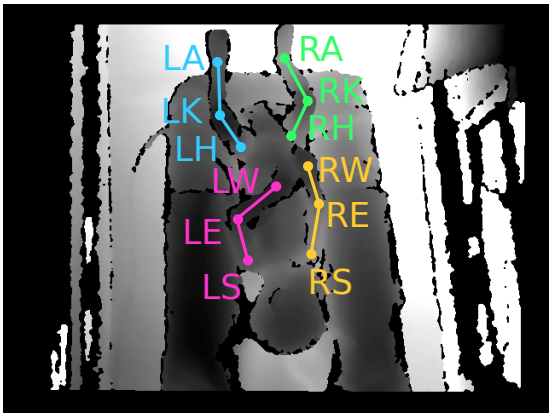


Figure 2.4: Preterm infant's joint model (including joints and joint-connections) superimposed on a sample depth frame. Inspired by clinical considerations, only limb joints are considered. LS and RS: left and right shoulder, LE and RE: left and right elbow, LW and RW: left and right wrist, LH and RH: left and right hip, LK and RK: left and right knee, LA and RA: left and right ankle.

is a publicly available three-year collection of preterm infants' depth videos acquired in the NICU of the "G. Salesi Hospital" (Ancona, Italy) with the set-up showed in Fig. 2.1.

The dataset is publicly available[2]: (i) to improve studies on the relationship between movement patterns and preterm-birth short and long-term complications, (ii) to support AI researchers that need raw data and corresponding annotation to train their models that may lead to significant enhancement in the field of preterm infants' movement monitoring from contactless measurements and (iii) to provide healthcare professionals working in NICUs with decision support in the evaluation of movement patterns. Two versions of the dataset were released: (i) the babyPose-v1 which accounted 16000 annotated depth frames from 16 preterm infants and (ii) the babyPose-v2 which adds to the previous version 11000 frames from 11 preterm infants (total number of annotated frames= 27000).

The NICU of the "G. Salesi" Hospital admits newborn infants born at term, born beyond term and preterm infants who require advanced care. Among infants born prematurely in the babyPose dataset, there can be: (i) infants born with extremely low gestational age and birth weight (i.e., among the 27, the preterm infant with the lowest gestational age was born at 24 weeks, weighs 850 grams and is 39 cm tall), (ii) infants with higher gestational age and weight (i.e., among the 27, the preterm infant with the highest gestational age was born at 37 weeks, weighs 3120 grams and is 37 cm tall) (iii) infants with congenital defects, metabolic and renal diseases, complex

[2] 10.5281/zenodo.3891404

surgical infants in the pre- and post-operative period infants with severe multi-system organ failure. Such a variability due to the clinical condition of the premature infant (and consequently the treatment to which the infant is subjected), gestational age, weight and height, is accompanied by intrinsic variability of the data typical of an acquisition conducted during the actual clinical practice. In this regard, the dataset has frames with external occlusions, frames in which the infants is not well positioned with respect to the camera and with variable pixel intensity. A proof of such data variability is shown in Figure 2.3.

2.2.1 Data annotation and ethical considerations

Infants' videos were acquired after obtaining the approval of: (i) the Ethics Committee of the "Ospedali Riuniti di Ancona", Italy (ID: Prot. 2019-399) and (ii) each infant's legal guardian who, aditionally, were asked to sign an informed consent of adhesion to the experimentation. The video recordings (each of which had a length of 180 s) were acquired with the Astra Mini S - Orbbec® camera, with a frame rate of 30 frames per second and image resolution of 640x480 pixels. Considering the average preterm infants' movement rate [50], in accordance with the SINC clinical partners, for each of the videos, 1 frame every 5 was extracted.

The proposed infant's model considers each of the 4 limbs as a set of 3 connected joints (i.e., wrist, elbow and shoulder for arms, and ankle, knee and hip for legs), as shown in Fig. 2.4. This choice is driven by the clinical consideration that monitoring legs and arms movement is of particular interest for evaluating preterm infants' cognitive and motor development [51, 52].

Limb-joints of each frame were manually annotated with a custom-built annotation tool, publicly available online[3]. For each of the infants, 1000 frames were annotated. The supervision of the annotation procedure by the clinicians was crucial especially in presence of challenging video frames (e.g., when the operators covered part of the joints during the actual clinical practice).

The following sections will present some DL methodologies to analyse data from the two versions of babyPose. The proposed studies fully respect and promote the values of freedom, autonomy, integrity and dignity of the person, social solidarity and justice, including fairness of access. The studies were carried out in compliance with the principles laid down in the Declaration of Helsinki, in accordance with the Guidelines for Good Clinical Practice.

[3] https://github.com/roccopietrini/pyPointAnnotator

2.3 Preterm infants' limb-pose estimation via spatio-temporal features analysis

The possibility of quantitatively assessing preterm infants' limb-movement via CNNs for video analysis, as highlighted in Sec. 2.1, has been unlocked by several researchers in literature [44, 46]. All the contributions have shown that, compared to more traditional analysis methods (e.g., based on wearable sensors posed over infants' limbs [36] or smart jumpsuits [38]), video-based monitoring systems are non-intrusive, low-cost, portable, and easy to use both in a clinical and domestic environment.

However, the approaches in [44, 46], besides analysing RGB data, only considered spatial features, without exploiting temporal information naturally encoded in video recordings [53]. A first attempt of including temporal information is proposed in [54], where RGB videos are processed by a semi-automatic algorithm for single-limb tracking. Motion-segmentation strategies based on particle filtering are implemented, which, however, relies on prior knowledge of limb trajectories. Such trajectories may have high variability among infants, especially when dealing with pathological infants, hampering the translation of the approach into the actual clinical practice. A possible alternative to exploit temporal information could be using 3D CNNs to directly extract spatio-temporal information from videos, which has already been shown to be robust in action recognition [55] as well as for surgical-tool detection [56].

Thus, guided by the research hypothesis that spatio-temporal features extracted from depth streams may boost performance with respect to spatial features alone, this section describes the first pipeline based on 3D CNNs for estimating preterm infants' limb pose from depth video recordings acquired in the actual clinical practice. The pipeline has a main detection CNN whose double role consists in: (i) roughly estimating the position of the joints and joint-connection and (ii) acting as a guidance for a regression CNN which refine the previously found predictions. The last step (i.e., the joint-linking step) deals with tracing the limbs skeleton from the regression CNN outcomes.

The innovative contributions of this research are summarized as follows:

1. Development of among the first DL pipeline for preterm infants' pose estimation from depth streams analysis. The pipeline exploits spatio-temporal features for automatic limb-joints and connections regression;

2. Validation of the approach from data collected in the actual clinical practice: a comprehensive study is conducted using 16 videos acquired in the actual clinical practice from 16 preterm infants (i.e., the babyPose dataset-v1) to experimentally investigate the research hypothesis.

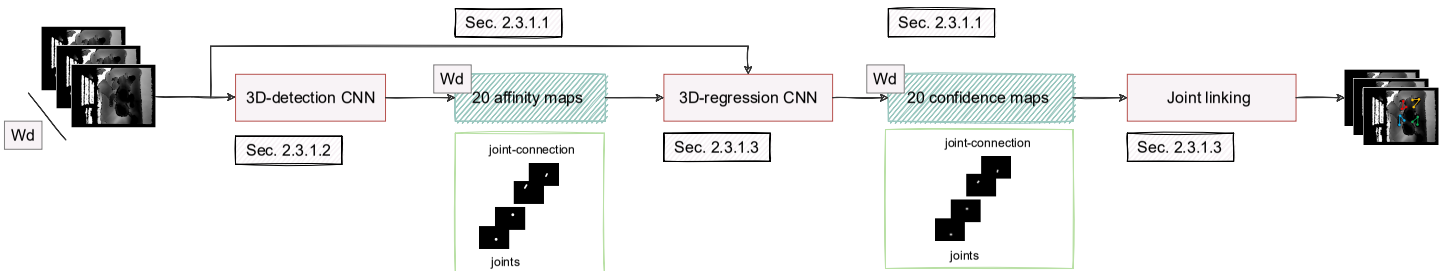


Figure 2.5: Workflow of the proposed pipeline to preterm infants' pose estimation with spatio-temporal features extracted from depth videos. The input consists of a temporal clip of W_d consecutive depth frames, which are processed by two convolutional neural networks (CNNs) to roughly detect joint and joint-connection (affinity maps) and refine joint and joint-connection detection (confidence maps), respectively.

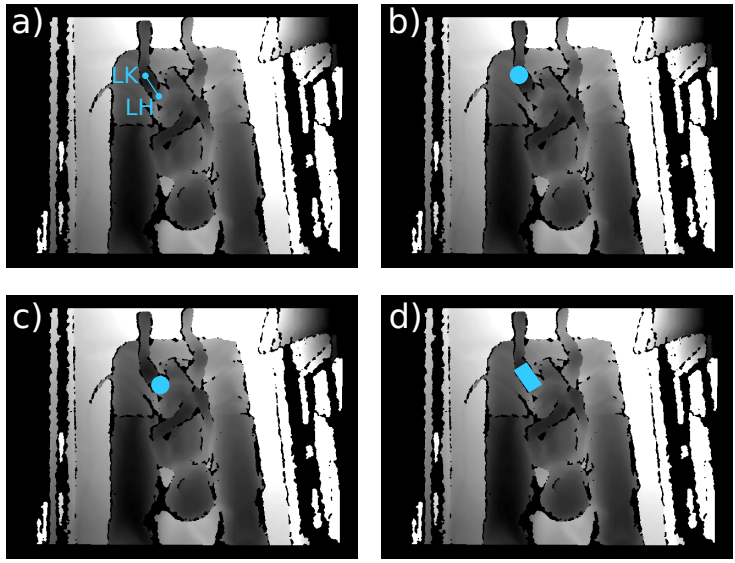


Figure 2.6: Ground-truth samples for the detection network. Samples are shown for (b) left knee (LK), (c) left hip (LH), and (d) their connection.

2.3.1 Methods

Figure 2.5 shows an overview of the workflow of the proposed spatio-temporal pipeline for preterm infants' pose estimation from depth streams.

Two consecutive CNNs were exploited, the former for detecting joints and joint connections, resulting in the so-called affinity maps, and the latter for precisely regressing the joint position, resulting in the confidence maps, by exploiting both the joint and joint-connection affinity maps, with the latter acting as guidance for joint linking. The joints belonging to the same limb are then connected using bipartite graph matching, following the model in Figure 2.4.

2.3.1.1 Data preparation

With the aim of extracting spatio-temporal features, temporal clips of depth frames were adopted. Following the approach presented in [56], a sliding window algorithm was implemented for building the clips: starting from the first video frame, an initial clip with a predefined number (W_d) of frames is selected and combined to generate a 4D datum of dimensions frame width (W) x frame height (H) x W_d x 1, where 1 refers to the depth channel. Then the window moves of W_s frames along the temporal direction and a new clip is selected.

To train the detection CNN, multiple binary-detection operations (considering each joint and joint-connection separately) were performed to solve possible ambiguities of multiple joints and joint connections that may cover the same frame portion (e.g.,

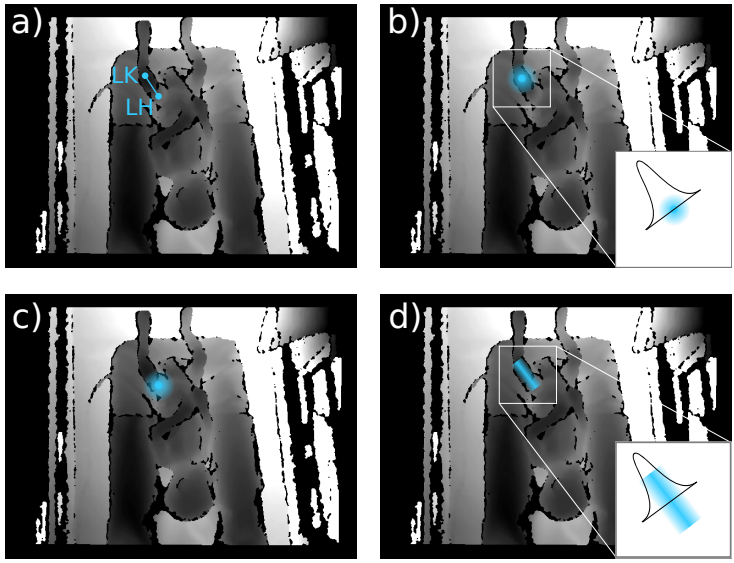


Figure 2.7: Ground-truth samples for the regression network. Samples are shown for (b) left knee (LK), (c) left hip (LH), and (d) their connection.

in case of self-occlusion). Hence, for each depth-video frame, 20 binary ground-truth affinity maps were constructed: 12 for joints and 8 for joint connections (instead of generating a single mask with 20 different annotations, which has been shown to perform less reliably [48]). Sample ground-truth maps are shown in Fig. 2.6. This results in a 4D datum of size W x H x W_d x 20. For each affinity map for joints, a region consisting of all pixels that lie in the circle of a given radius (r_d) centered at the joint center was considered.

A similar approach is used to generate the ground-truth affinity map for the connections. In this case, the ground-truth is the rectangular region with thickness r_d and centrally aligned with the joint-connection line. The regression CNN is fed by stacking the depth temporal clip and the corresponding affinity maps obtained from the detection network. Thus, the regression input is a 4D datum of dimension W x H x W_d x 21 (i.e., 1 depth channel + 12 joints + 8 connections). The regression network is trained with W_d x 20 ground-truth confidence maps of size W x H (Fig. 2.7). For every joint in each depth frame, a region of interest consisting of all pixels that lie in the circle with radius r centered at the joint center was considered. In this case, instead of binary masking the circle area as for the detection CNN, a Gaussian distribution with standard deviation (σ) equal to 3*r and centered at the joint center was implemented. A similar approach is used to generate the ground-truth confidence maps for the joint connections. In this case, the ground-truth map is the rectangular region with thickness r and centrally aligned with the joint-connection line. Pixel values in the mask are 1-D Gaussian distributed (σ = 3*r) along the connection direction.

2.3.1.2 3D detection convolutional neural network

The detection CNN (Figure 2.8) is inspired by the classic encoder (Enc)-decoder (Dec) architecture of U-Net [57], which is however implemented as a two-branch architecture for processing joints and joint connections separately. In fact, using a two-branch architecture has been shown to provide higher detection performance for 2D architecture [48, 56]. To incorporate the spatio-temporal information encoded in infants' depth streams, 3D CNN kernels were used. The 3D convolution allows the kernel to move along the 3 input dimensions to process multiple frames at the same time, preserving and processing temporal information through the network.

The detection CNN starts with an input layer and a common-branch convolutional layer (with stride = 1 and kernel size = 3x3x3 pixels), and is followed by 8 blocks. Each block is first divided in two branches (for joints and connections). In each branch, two convolutions are performed: the former with kernel size = 2x2x2 and stride 2x2x1, while the latter with kernel size = 3x3x3 and stride 1x1x1. It is worth noting that kernel stride is equal to 1 in the temporal dimension as to avoid deteriorating meaningful temporal information. The outputs of the two branches in a block are then concatenated in a single output, prior entering the next block. In each block of the Enc path, the number of channels is doubled. Batch normalization and activation with the rectified linear unit (ReLu) are performed after each convolution.

The architecture of the decoder path is symmetric to the Enc one and ends with an output layer with W_dx20 channels (12 for joints and 8 for connections) activated with the sigmoid function.

2.3.1.3 3D regression convolutional neural network and joint linking

The necessity of using a regression CNN for the addressed task comes from considerations of previous work [58], which showed that directly regressing joint position from an input frame is highly non linear. The regression network, instead, produces W_dx20 stacked confidence maps (12 for joints and 8 for connections). Each map has the same size of the input depth clip (i.e., WxH).

Also in this case, 3D convolution is performed to process spatio-temporal features. The network consists of five layers of 3x3x3 convolutions (Figure 2.9). Kernel stride is always set to 1, to preserve the spatio-temporal resolution. In the first 3 layers, the number of activations is doubled, ranging from 64 to 256. The number of activations is then kept constant for the last two layers. Batch normalization and ReLu-activation are performed after each 3D convolution.

The last step of the limb pose-estimation task deals with linking subsequent joints for each of the infants' limb, which is done on depth images, individually. First, joint candidates were identified from the output joint-confidence maps using non-maximum suppression. This is an algorithm commonly used in computer vision when redundant candidates are present [59]. Once joint candidates are identified, they are linked

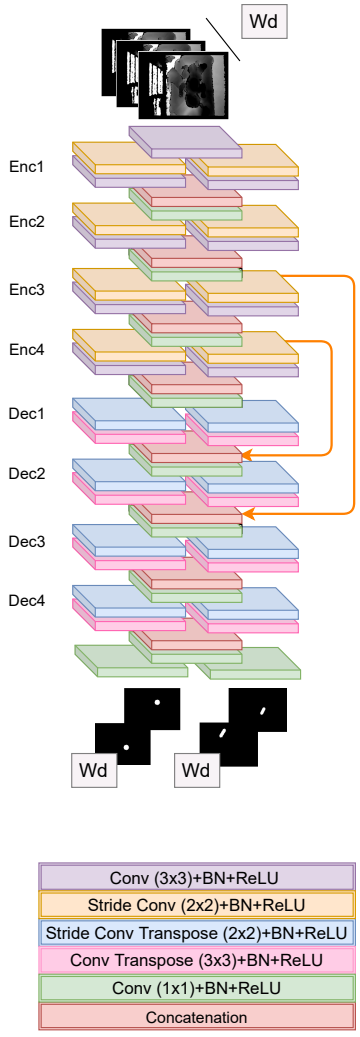


Figure 2.8: 3D detection convolutional neural network (CNN) for preterm infants' joints and joint-connections position estimation. It takes in input a stack of W_d=3 depth frames and outputs 60 (W_d x 20) affinity maps. The different colours of the convolutional blocks are coded in the legend on the bottom of the figure.

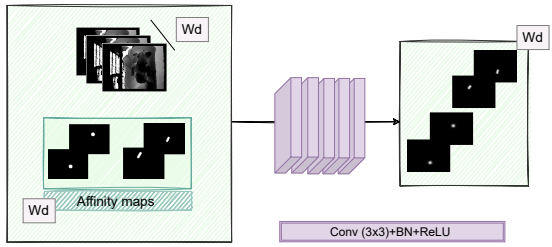


Figure 2.9: 3D regression convolutional neural network (CNN) for preterm infants' joints and joint-connections position estimation. This network takes in input the W_d=3 depth frames and the output of the detection network (i.e., 60 affinity maps) and outputs 60 (W_d x 20) confidence maps.

exploiting the joint-connection confidence maps via a bipartite graph matching approach, which consists of: (i) computing the integral value along the line connected two candidates on the joint-connection confidence map and (ii) choosing the two winning candidates as those guaranteeing the highest integral value.

2.3.2 Experimental protocol

2.3.2.1 Dataset

As described in Sec. 2.2, the babyPose-v1 was used for the experiments. It consisted of 16 depth videos (length = 180 s) of 16 preterm infants. The infants were identified by clinicians in the NICU among those who were spontaneously breathing.

For each of the videos 1000 frames were manually annotated. Then, these 1000 frames were split into training and testing data: 750 frames were used for training purpose and the remaining ones (250 frames) to test the network. This resulted in a training set of 12000 samples (16 infants x 750 frames) and a testing set of 4000 samples (16 infants x 250 frames). From the 12000 training samples, 200 frames for each infant were kept as validation set, for a total of 3200 frames.

2.3.2.2 Training settings

All frames were resized to 128x96 pixels in order to smooth noise and reduce both training time and memory requirements. Mean intensity was removed from each frame.

To build the ground-truth masks, r_d equal to 6 pixels was selected, as to completely overlay the joints. The W_s was set to 2 for training and 0 for testing, while W_d was set to 3. This way, a temporal clip was 0.5 s long.

For training the 3D detection and 3D regression CNNs, an initial learning rate of 0.01 with a learning decay of 10% every 10 epochs was set while the momentum was

equal to 0.98. A batch size of 8 and a number of epochs equal to 100 were set for training the CNNs.

The 3D detection CNN is trained using the adaptive moment estimation (Adam) as optimizer and the per-pixel binary cross-entropy (L_{CE}), adapted for multiple 3D map training, as loss function:

$$L_{CE} = \frac{1}{W_d(J+C)\Omega} \sum_{t=1}^{W_d} \sum_{k=1}^{J+C} \sum_{\mathbf{x}\in\Omega} [p_{t,k}(\mathbf{x}) log(\tilde{p}_{t,k}(\mathbf{x})) + (1-p_{t,k}(\mathbf{x})) log(1-\tilde{p}_{t,k}(\mathbf{x}))] \quad (2.1)$$

where $p_{t,k}(\mathbf{x})$ and $\tilde{p}_{t,k}(\mathbf{x})$ are the ground-truth affinity maps and the corresponding output at pixel location $\mathbf{x}$ in the depth-frame domain (Ω) of channel k for temporal frame t, J=12 and C=8 are the number of joints and joint connections, respectively.

The regression network is trained with the stochastic gradient descent as optimizer using the mean squared error (L_{MSE}), adapted for multiple 3D map training, as loss function:

$$L_{MSE} = \frac{1}{(J+C)\Omega} \sum_{t=1}^{W_d} \sum_{k=1}^{J+C} \sum_{\mathbf{x}\in\Omega} [h_{t,k}(\mathbf{x}) - \tilde{h}_{t,k}(\mathbf{x})] \quad (2.2)$$

where $h_{t,k}(\mathbf{x})$ and $\tilde{h}_{t,k}(\mathbf{x})$ are the ground truth and the predicted value at pixel location $\mathbf{x}$ of the k^{th} channel for temporal frame t, respectively.

The best model was selected as the one that maximized the detection accuracy (Acc) and minimized the mean absolute error on the validation set, for the 3D detection and 3D regression networks, respectively. All the analyses were performed using Keras[4] framework on a Intel® Xeon® Silver 4214 CPU @ 2.20GHz with 230 GB of RAM and a NVIDIA® RTX 2080 8 GB RAM.

2.3.2.3 Ablation study and comparison with the state of the art

The performance of the proposed pipeline was compared with that of the same pipeline for spatial features analysis only. The 2D pipeline is inspired by [48] and uses the same architectures presented in Figure 2.8 and Figure 2.9, but with 2D spatial convolution for the analysis of single-depth frames. The 2D-convolution-based CNNs were identified as 2D detection and 2D regression CNN, respectively.

The proposed approach was further compared against Stacked Hourglass [49] and Convolutional Pose Machine [60], which are among the most successful and well-known approaches for human-pose estimation. For these comparisons, the corresponding architectures, originally designed for RGB images, were modified to allow depth-image processing. For all these architectures, the same training settings described in Sec. 2.3.2.2 were implemented.

[4] https://keras.io/

For the ablation study, inspired by [61], the performance of the proposed pipeline was compared with the detection-only and regression-only architectures. Both were implemented in a spatio-temporal fashion (i.e., with 3D convolutions). For the detection-only model, the affinity maps were used to directly estimate limb pose with the bipartite graph matching strategy (Sec. 2.3.1.3). The regression-only model was fed with the depth clips and trained with the confidence-map ground truth. The output was then used to estimate joint pose with bipartite matching.

2.3.2.4 Performance metrics

To measure the performance of the detection network, as suggested in [56], both the Dice similarity coefficient (*DSC*) and recall (*Rec*) were computed, respectively. These are defined as follows:

$$DSC = \frac{2 \times TP}{2 \times TP + FP + FN} \tag{2.3}$$

$$Rec = \frac{TP}{TP + FN} \tag{2.4}$$

where TP and FP are the true joint and background pixels detected as joints, respectively, while FN refers to joint pixels that are detected as background. The same applied to joint-connections.

To evaluate the overall pose estimation, the root mean square distance (*RMSD*) for each infants' limb was assessed. For both the detection and regression network the testing time was reported.

Two-sided t-test with significance level (α) = 0.05 was used to evaluate if significative differences were present between the 2D and 3D pipeline in estimating limbs pose.

2.3.3 Results

The descriptive statistics of *Rec* and *DSC* for the detection CNN are reported in Table 2.1. Figure 2.10 shows the *Rec* boxplots for joints. Results are also shown for the corresponding 2D implementation. The highest median *DSC* (0.94, inter-quartile range (IQR) = 0.05) among all joints was obtained with the 3D CNN. The same was observed for the *Rec*, with a median value among all joints of 0.90, and IQR of 0.09. Note that, in the case yielding the least accurate result, which corresponds to the RH joint, the *Rec* still achieved 0.88, whereas for the 2D detection network the lowest *Rec* was 0.73. The same behaviour (Table 2.2 and Fig. 2.10) was observed when considering the joint-connection detection performance, with median *DSC* = 0.93 (IQR = 0.06) and median *Rec* = 0.90 (IQR = 0.11) among all connections.

The performance comparison in terms of *RMSD* of the different models presented in Sec. 2.3.2.3 is summarized in Table 2.3. The highest performance (i.e., the lowest

Table 2.1: Joint-detection performance in terms of median Dice similarity coefficient (*DSC*) and recall (*Rec*). Inter-quartile range is reported in brackets. The metrics are reported separately for each joint. For joint acronyms, refer to the joint-pose model in Fig. 2.4.

	Right arm			Left arm			Right leg			Left leg		
	RW	RE	RS	LS	LE	LW	RA	RK	RH	LH	LK	LA
DSC												
2D detection CNN	0.84 (0.11)	0.87 (0.10)	0.86 (0.09)	0.87 (0.09)	0.85 (0.08)	0.86 (0.09)	0.86 (0.10)	0.87 (0.07)	0.84 (0.08)	0.85 (0.09)	0.86 (0.07)	0.87 (0.07)
3D detection CNN	**0.93** (0.06)	**0.94** (0.06)	**0.94** (0.07)	**0.94** (0.08)	**0.94** (0.06)	**0.94** (0.06)	**0.93** (0.06)	**0.94** (0.05)	**0.93** (0.06)	**0.93** (0.06)	**0.94** (0.05)	**0.93** (0.06)
Rec												
2D detection CNN	0.73 (0.15)	0.77 (0.15)	0.76 (0.11)	0.78 (0.13)	0.73 (0.11)	0.76 (0.14)	0.77 (0.15)	0.78 (0.11)	0.73 (0.11)	0.74 (0.12)	0.76 (0.12)	0.78 (0.10)
3D detection CNN	**0.89** (0.11)	**0.90** (0.10)	**0.91** (0.11)	**0.91** (0.12)	**0.90** (0.09)	**0.90** (0.09)	**0.89** (0.09)	**0.90** (0.09)	**0.88** (0.09)	**0.89** (0.09)	**0.92** (0.07)	**0.89** (0.11)

Table 2.2: Joint-connection detection performance in terms of median Dice similarity coefficient (*DSC*) and recall (*Rec*). Inter-quartile range is reported in brackets. The metrics are reported separately for each joint connection. For joint acronyms, refer to the joint-pose model in Fig. 2.4.

	Right arm		Left arm		Right leg		Left leg	
	RW-RE	RE-RS	LS-LE	LE-LW	RA-RK	RK-RH	LH-LK	LK-LA
	DSC							
2D detection CNN	0.89 (0.08)	0.90 (0.08)	0.89 (0.07)	0.88 (0.08)	0.90 (0.06)	0.88 (0.08)	0.90 (0.07)	0.91 (0.06)
3D detection CNN	**0.93** (0.06)	**0.93** (0.08)	**0.94** (0.08)	**0.94** (0.06)	**0.93** (0.05)	**0.93** (0.06)	**0.94** (0.06)	**0.94** (0.06)
	Rec							
2D detection CNN	0.81 (0.12)	0.84 (0.13)	0.81 (0.12)	0.81 (0.12)	0.85 (0.09)	0.80 (0.12)	0.85 (0.12)	0.85 (0.09)
3D detection CNN	**0.90** (0.10)	**0.89** (0.15)	**0.92** (0.13)	**0.90** (0.10)	**0.90** (0.08)	**0.88** (0.09)	**0.91** (0.09)	**0.90** (0.10)

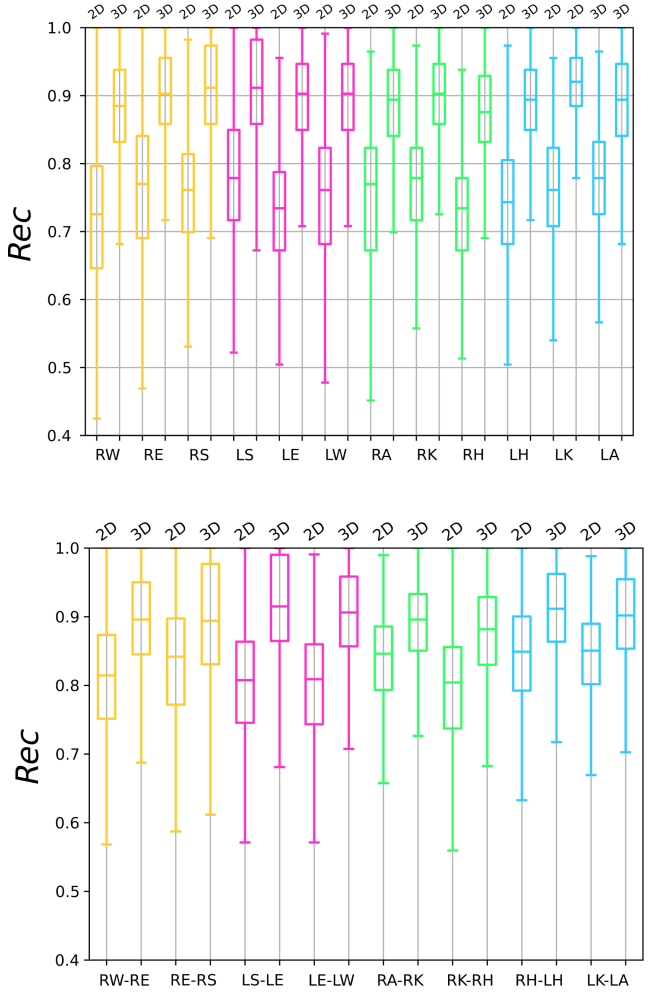


Figure 2.10: Boxplots of the recall (*Rec*) for joint (top) and joint-connection (bottom) detection achieved with the proposed 3D pipeline. Results of its akin 2D are shown for comparison, too. For colors and acronyms, refer to the joint model in Fig. 2.4.

Table 2.3: Limb-pose estimation performance in terms of median root mean square distance (*RMSD*), with interquartile range in brackets, computed with respect to the ground-truth pose. The *RMSD* is reported for each limb, separately. Results are reported for the 2D and 3D pipeline, as well as for the 3D detection-only, 3D regression-only and state-of-the-art architectures.

	Right arm	Left arm	Right leg	Left leg
	RMSD			
2D pipeline	11.73 (3.58)	10.54 (4.97)	11.03 (5.78)	11.50 (4.21)
Detection-only network	15.09 (3.80)	15.60 (3.87)	15.09 (3.41)	14.91 (3.49)
Regression-only network	12.39 (2.18)	11.73 (3.25)	11.95 (4.60)	12.17 (2.47)
Stacked Hourglass	13.01 (4.12)	11.95 (4.60)	11.27 (5.32)	11.95 (3.58)
Convolutional Pose Machine	12.17 (4.52)	11.73 (3.65)	11.27 (4.61)	11.95 (3.44)
3D pipeline	**9.76** (4.60)	**9.29** (5.89)	**8.90** (5.64)	**9.20** (3.99)

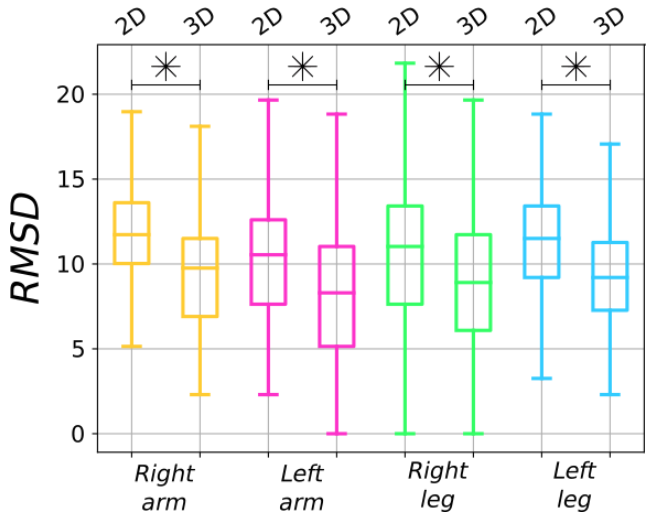


Figure 2.11: Boxplots of the root mean squared distance (*RMSD*) computed for the four limbs separately. Boxplots are shown for the 2D and 3D pipeline. Asterisks highlight significant differences.

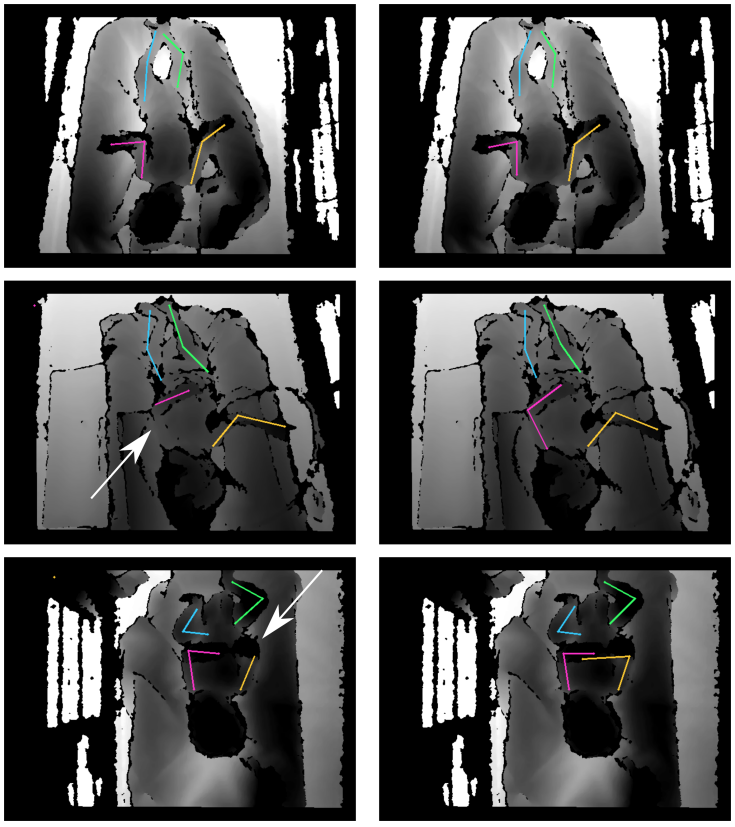


Figure 2.12: Sample qualitative results for pose estimation obtained with the 2D (left) and 3D (right) pipeline. White arrows highlight estimation errors, mainly due to homogeneous image intensity.

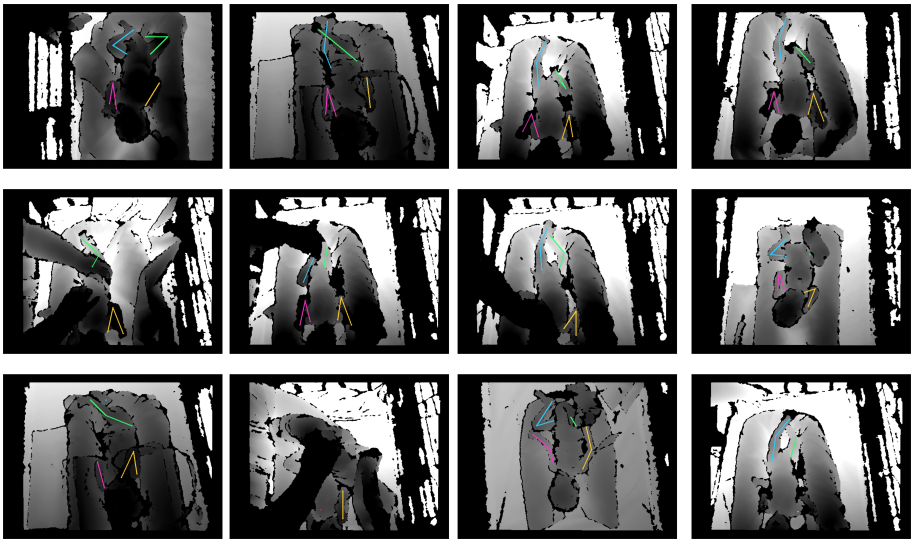


Figure 2.13: Sample qualitative results for challenging cases. First row: one joint was not detected due to auto-occlusion (from left to right: right shoulder, right shoulder, right hip, right hip). Second row: one or more joints were not detected due to external occlusion (from left to right: joint of the left limbs, right ankle, left arm - due to healthcare operator hand presence, and right knee and ankle - due to plaster). Last row: image noise and intensity inhomogeneities prevented joint detection.

RMSD) was achieved by the 3D pipeline, with a median value of 9.06 pixels (IQR = 5.12) among the four limbs. The best performance was achieved for the right leg (median = 8.90 pixels, IQR = 5.64 pixels). The overall computational time for the 3D pipeline was 0.06 s per image on average. The 2D pipeline always showed lower performance, with the best and worst *RMSD* equal to 10.54 (left arm) and 11.73 (right arm) pixels, respectively (median among the four limbs = 11.27 with IQR = 4.59). The overall statistics are shown in Fig. 2.11. The results all differed significantly (p-value $< \alpha$) from those obtained with the 3D pipeline. Stacked Hourglass and Convolutional Pose Machine got a median *RMSD* of 11.95 and 11.84 pixels. The detection-only and regression-only networks showed the lowest performance, with a median *RMSD* equal to 15.09 pixels and 12.06 pixels, respectively.

In Fig. 2.12, qualitative results for infants' pose estimation are shown both for the 2D pipeline (on the left side) and the 3D one (on the right side). The white arrows highlight errors in pose estimation made by the 2D pipeline. Results of the 3D pipeline for challenging cases are shown in Fig. 2.13. The first row shows samples in which one joint was not detected due to auto-occlusion. Joints were also not detected when external occlusion occurred (second row), due to the interaction of the healthcare-operator with the infant or to the presence of plaster. The proposed pipeline was unable to correctly estimate limb-pose also when image noise and intensity inhomogeneities (e.g., due to rapid infants' movement) were present (third row). At the same time, however, other joints in the image were correctly estimated thanks to the joint-map parallel processing.

2.3.4 Discussion

Monitoring preterm infants' limb is crucial for assessing infant's health status and early detecting cognitive and motor disorders. However, when surveying the clinical literature, there is a lack of documented quantitative parameters on the topic. This is mainly due to the drawbacks of current monitoring techniques, which rely on qualitative visual judgment of clinicians at the crib side in NICUs. A possible, straightforward, solution may be to exploit contact sensors (such as accelerometers). Nonetheless, in NICUs, using additional hardware may contribute significantly to infants' stress, discomfort and pain and, from the healthcare operators' point of view, may hinder the actual clinical practice. To overcome all these issues, researchers seek for new reliable and unobtrusive monitoring alternatives, which are mostly based on video analysis. This section described a novel pipeline for non-invasive monitoring of preterm infants' limbs providing an innovative approach for limb-pose estimation from spatio-temporal features extracted from depth streams. The choice of processing depth videos (over RGB ones) was driven by the necessity to fully protect the ward privacy. While the rationale behind the use of 3D convolutions instead of networks inherently capable of processing temporal information (e.g., long-short term memory

networks) lies on the fact that preterm infants movements may be sporadic and of short duration.

The DL pipeline was validated on the babyPose dataset-v1, whose video recordings, acquired in the actual clinical practice, presented several challenges such as: presence of homogeneous areas with similar or at least continuous intensity, self- or external occlusions and different pose of the camera with respect to the infants.

The proposed 3D detection network achieved encouraging results as shown in Figure 2.10 and reported in Table 2.1, with a median *DSC* of 0.94 and 0.93 for joints and joint-connections, respectively, overcoming the detection CNN based on spatial features only (i.e., 2D detection CNN). The network performed comparably when detecting all joints and joint-connections as shown by the IQRs in Table 2.1, reflecting the CNN ability of processing in parallel the different joint and joint-connection affinity maps.

The 3D pipeline achieved improved performance (Table 2.3) in estimating infants' pose for all limbs (median $RMSD = 9.06$ pixels) when compared with its akin with 2D convolutions (median $RMSD = 11.27$ pixels). These results suggest that exploiting temporal information improved network generalization ability even in presence of intensity homogeneity and noisy background, typical of depth images. These considerations are visible in Fig. 2.12, where the 2D pipeline failed in detecting joints that lay in portions of the image with homogeneous intensity.

Predictions of the pose estimation were computed also for the detection- (median $RMSD = 15.09$ pixels) and the regression-only networks (median $RMSD = 12.06$ pixels). Despite the complexity of regressing joint and joint-connection confidence maps from depth image clips only, the regression-only network achieved better results when compared to the detection-only network. The lower performance of the detection-only network may be due to the complexity in localizing joint candidates from ground-truth binary masks, where all pixels have the same weight (Fig. 2.6). It is worth noting that spatio-temporal features were tested for a detection-only task in [56] (even though for surgical instrument joints in laparoscopic video). The proposed work, however, moved forward to test joint estimation by combining the detection network with bipartite matching, and comparing the achieved results with the full 3D pipeline (i.e., 3D detection and 3D regression). Despite the integration of the temporal information, both the detection-only and regression-only network achieved lower outcomes with respect to the full 2D pipeline. Hence, the regression-only model was barely capable of predicting the location of joints without any guidance. Regression is empirically too localized (i.e., it supports small spatial context) and the process of regressing from original input image to joint location directly is challenging. By combining detection and regression, the detection module acted as structural guidance for the regression module by providing spatial contextual information between joints, and facilitating the joints localization.

Stacked Hourglass and Convolutional Pose Machine achieved lower performance

when compared to our 3D pipeline. This might be attributed to the fact that both Stacked Hourglass and Convolutional Pose Machine are designed to process spatial features only. Nonetheless, the 2D pipeline, which also works with spatial feature only, overcame both Stacked Hourglass and Convolutional Pose Machine. This result seems to confirm that the rough detection of limb joints by the detection network facilitates the regression network in regressing joint position accurately, as highlighted in [48]. In fact, Stacked Hourglass and Convolutional Pose Machine achieved better *RMSD* values when compared to the regression-only network. Hence, the benefits brought by the introduction of 3D kernels in the regression-only network are counterbalanced by the multi-scale nature of the state-of-the-art networks, which capture both global and local information.

A straightforward limitation of this work may be seen in the estimation of occluded joints (both in case of auto and external occlusion), as shown in Fig. 2.13 (first and second rows). At the same time, the two-branch architecture with multiple maps allowed to detect the other (not-occluded) joints in the image. This issue could be attenuated with recent strategies proposed in the literature for long-term tracking [62] and confidence estimation [63]. Modeling infant's limbs through anthropometric measures (such as limb length - already acquired in the actual clinical practice) could also help in attenuating the occlusion issue. This would probably also make the 3D pipeline able to tackle noisy image portions, which may be present due to sudden movement of infants or healthcare operators (Fig. 2.13, last row). A limitation of the proposed work could be seen is the relatively limited number of testing frames (4000), which is due to the lack of available annotated dataset online. This, however, further motivated the decision to release the dataset to the scientific community.

2.4 Dense-atrous spatial-convolutional blocks to estimate preterms' limb-pose

In Sec. 2.3, the potential of a 3D pipeline for estimating preterm infants' limb pose was proved. The pipeline processed depth streams (i.e., clips with 3 subsequent depth frames) via two spatio-temporal CNNs (i.e., 3D detection CNN and 3D regression CNN), improving the experimental results on the babyPose-v1 dataset with respect on its akin which processes depth frames (i.e., 2D detection CNN and 2D regression CNN). Despite its robustness, the 3D pipeline is computationally expensive as it relies on 3D convolutions to process spatio-temporal features, raising concerns relevant to real-time and on-the-edge processing. As claimed in a recent meta-analysis [47], despite the promising advancements in the literature, further studies are still needed to (i) tackle the high-image variability and (ii) limit the computational resources needed, with a view to design embedding solutions easily deployable also in a domestic environment.

Guided by these premises, this research proposes an improved CNN-based pipeline for estimating preterm infants' limb-pose from the analysis of depth frames only. Inspired by literature in closer fields [1, 64], the pipeline takes advantage of the generalization power of the dense blocks [65] and atrous separable convolutions [66]. The former strengthen features propagation throughout network layers while the latter enables multiple scale information processing while keeping the resolution intact. As proved in [1], this innovation has the potentiality of improving a CNN ability to finely exploit semantic information by gathering fine details from a context-aware analysis, while attaining low number of network parameters.

The improvement of the research here presented is twofold: it boosts the performance of the 2D-pipeline for pose estimation drastically reducing the computation cost naturally induced by the 3D convolutions implemented by the 3D pipeline.

The innovative contributions of the research are summarized as follows:

1. Improving the performance of the 2D pipeline leveraging a DeA-based convolutional pathway while relying on spatial features only. The pathway couples the dense blocks [65] and atrous separable convolutions [66] to improve the performance of the pipeline on challenging depth images (e.g., images with low intensity or with few joints due to the presence of several occlusions) [1];

2. Validation of the proposed approach on the expanded version of babyPose-v1. The new dataset (i.e., babyPose-v2 dataset) counts 27 videos from 27 preterm infants. All the videos were acquired in the NICU of the "G. Salesi" Hospital in Ancona still from infants' who spontaneously breathing;

3. Introducing a sustainability plan aimed at finding a solution that provides both effectiveness and efficiency. The proposed pipeline has undergone an architec-

tural variation process aimed at finding its own optimized variant in terms of number of network parameters. This validation is crucial with a view to deploy the proposed pipeline in embedding systems for on-the-edge computation.

2.4.1 Methods

Figure 2.14 shows the workflow of the proposed pipeline to preterm infants' pose estimation. As in Sec. 2.3, the pipeline leverages two consecutive CNNs: the first CNN, called DeA-detection network, plays the key role of roughly detecting the joint- and joint-connection positions from the depth frame analysis. The second CNN, the DeA-regression network, leverages on the output of the first CNN to precisely retrieve the joints and joint-connections position. The last step, namely the join-linking step, exploits the outcome of the DeA-regression network to link each-limb consecutive joints.

2.4.1.1 DeA-detection convolutional neural network

The overall structure of the DeA-detection network is shown in Fig. 2.15. The network is fed with a depth image and outputs 20 affinity maps. It shares the same Enc-Dec structure of the 2D and 3D detection CNN, with 8 double-branch convolutional blocks: 4 for the Enc path (which performs downsampling) and 4 for the Dec path (which performs upsampling). In both the Enc and Dec path, each block is divided in two branches allowing for parallel processing joints and joint-connections. Each branch implements two convolutions (in the Enc path) and deconvolutions (in the Dec path). The output of each single-branch is concatenated to enter in a single convolutional block prior being newly forked.

Inspired by [1], The DeA-detection network implements a flowing pathway between the second Enc block (Enc2) and Dec block (Dec2). This pathway combines information from the classical long-skip connections [57] and the DeA pathway that couples atrous separable convolutions [66] and dense layers [65].

This architectural design was guided by the consideration that classical long-skip connections between the Enc and Dec path are crucial to recover the spatial information lost during downsampling operations. However, coupling of features from particularly shallow (Enc) and deep (Dec) layers via long-skip connections has two main drawbacks: (i) fails in localizing joints in challenging images (i.e., images with noise, barely visible joints and several joints occlusion) [1] (ii) induces the semantic gap due to multilevel features aggregation [67]. To bridge these issues, the flowing pathway fuses feature maps from long skip-connection with those of DeA-pathway.

As shown in Fig. 2.15, DeA pathway implements dense connections between atrous-blocks. This potentially enables smoothly information flow. Throughout dense connections, each atrous block gets inputs from all preceding blocks and distribute its

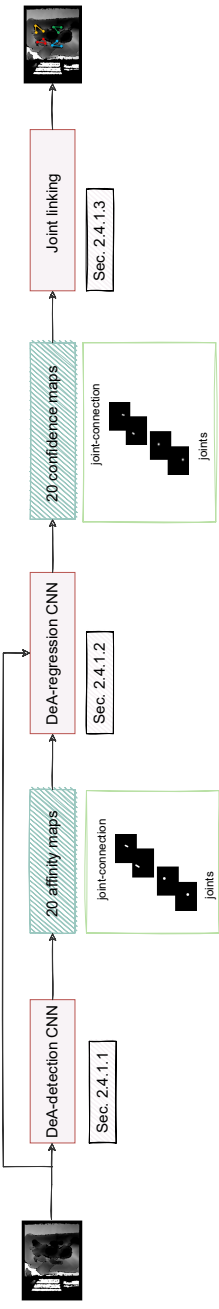


Figure 2.14: Workflow of the proposed pipeline to preterm infants' pose estimation with spatial features extracted from depth frames. The input consists of a depth frame processed by two convolutional neural networks (CNNs) (i.e., Dea-detection CNN and DeA-regression CNN) to roughly detect joint and joint-connection (affinity maps) and refine joint and joint-connection detection (confidence maps), respectively.

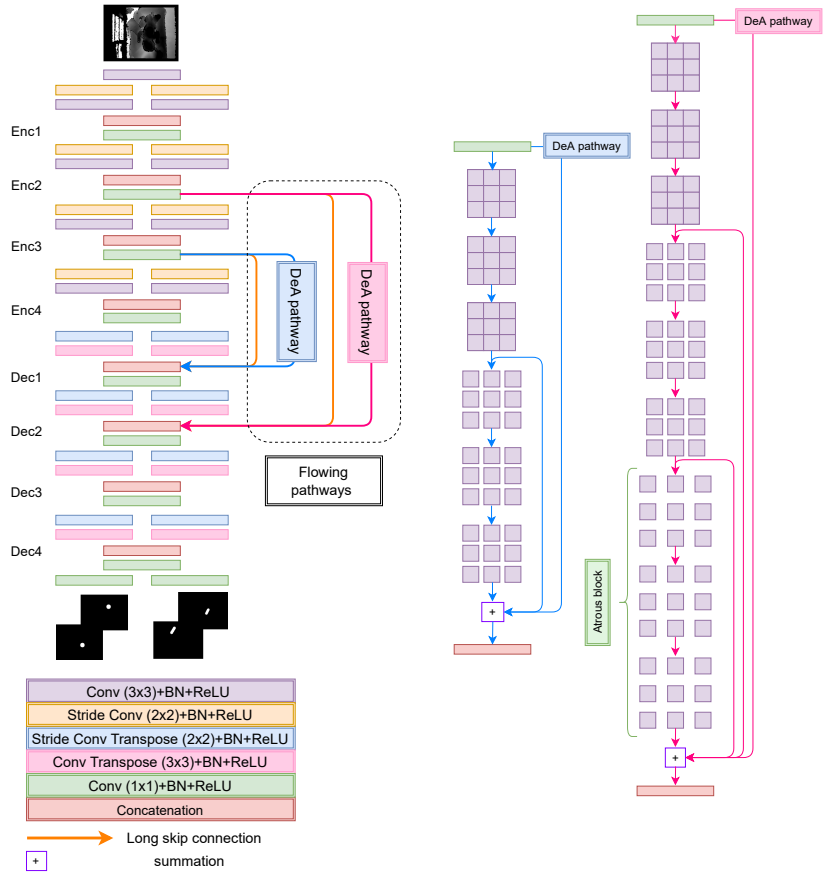


Figure 2.15: Graphical representation of the dense-atrous-based (DeA)-detection network. The body of the network consists of 8 convolutional layers (4 layer for the encoding (Enc) path and 4 layers for the decoding (Dec) path). Each coloured block in a convolutional layer implements: (i) the convolution operation, differing for kernel sizes (3x3 or 2x2 or 1x1), (ii) the batch normalization (BN) and (iii) the activation function (Rectified Linear Unit (ReLU)). Inspired by [1], the flowing pathway between Enc2-Dec2 couples the classical long-skip connection (orange) and the DeA pathway (pink). The detail of the DeA pathway between Enc2-Dec2 (pink) is shown on the far right of the figure. The colour-coded legend is located in the lower left corner.

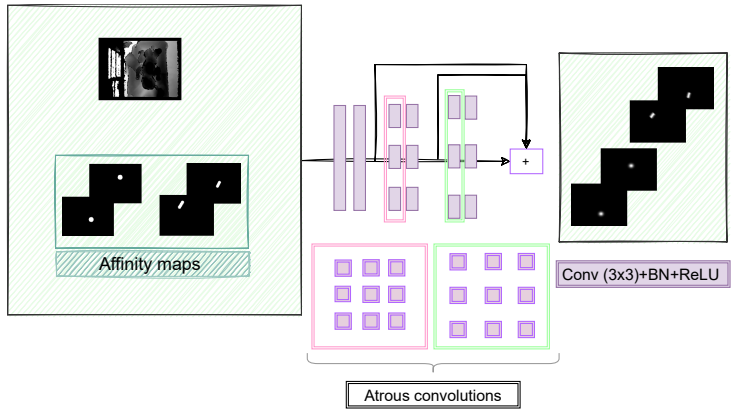


Figure 2.16: Graphical representation of the dense-atrous-based (DeA)-regression network. The input to the network consists of a stack with the input image and 20 affinity maps, which are produced from the DeA detection network, while the output consists of 20 confidence masks. The body of the network consists of 6 convolutional layers. Each violet block in a convolutional layer implements: (i) the convolution operation (kernel size= 3x3), (ii) the batch normalization (BN) and (iii) the activation function (Rectified Linear Unit (ReLU)).

own feature-maps to the next ones. Atrous blocks implement atrous separable convolutions. These convolutions deliver a wider field of view by dilating consecutive kernel values of a factor d. Coupling atrous blocks differing for the d factor in a dense fashion allows deeper layers to naturally gather multiple scale information together while keeping network parameters low. Herein, the atrous blocks consist of 3 subsequent convolutions sharing the same kernel size and dilation rate (d). In the proposed implementation the flowing-pathway has a DeA-pathway with 3 atrous blocks (d=1,2,3).

Batch normalization and activation with the ReLu are performed after each convolutional and deconvolutional block.

As for the 2D and 3D detection to train the DeA-detection CNN 20 affinity maps were prepared. For each joint, all the pixels lying within a circle of radius r centered at the manually annotated joint site were constructed. Similarly, the joint-connection ground-truth, is the rectangular region of thickness r and centrally aligned with respect to the line linking two consecutive joints.

2.4.1.2 DeA-regression convolutional neural network and limb-skeleton tracing

The goal of the pipeline is to predict the location of joints and connections between them to trace preterm infants' limb skeleton. As emerged in Sec. 2.3, directly deriving

a pair of coordinates describing the position of the joints and the joint-connections in space, is an hard and highly non-linear task [48]. Thus a pipeline with a DeA detection network was exploited to crudely track the position of the joints and joint-connections and a DeA regression network to refine the predictions of the previous CNN. The DeA regression network implements dense blocks of atrous convolutions. This choice derives from the hypothesis that collective knowledge of contextual information, captured via DeA layers, may guide the network on where to focus in an image enriching its ability to correctly regress joints location [58].

The input to the DeA regression network (Fig. 2.16) is a multichannel representation consisting of a stack with the affinity maps, in output from the DeA detection network, and the initial depth image. The output of the DeA regression CNN consists in 20 confidence maps (12 for joints and 8 for joint-connections) that provide a more precise location of joints and their connections. This mapping operation that redistributes the binary input maps in a Gaussian manner can be performed via a non-complex single-branch architecture designed not to burden the overall pipeline [58].

The DeA regression network has 3 atrous blocks with two subsequent convolutions sharing the same d factor (d=1,2,3) and an output layer. As for the DeA detection network, information thought the network flow in a dense fashion. Batch normalization and activation with ReLu are performed after each convolution.

As for the 2D and 3D regression CNNs, Gaussian-distributed masks (or confidence maps) are used to train the DeA-regression CNN. For each joint, a region consisting of all pixels laying in the circle of radius r centered at the manual annotation site were considered. Such region is the Gaussian distributed version of the binary mask, with Gaussian standard deviation equal to $3r$. For joint connections, the Gaussian-distributed version of the joint-connection affinity masks are built along the connection direction with a standard deviation equal to $3r$.

As for the 3D and 2D pipeline the last step consists in linking subsequent joints to trace the skeleton of each infants' limb. This step exploits the output predictions from the DeA regression network to connect subsequent joints and trace each limb-segment. Non-maximum suppression was performed to select joint-candidates from the joint-confidence map. Then a bipartite matching approach was implemented to choose, among the joint candidates, the two subsequent joints to trace the limb-skeleton.

2.4.2 Experimental Protocol

2.4.2.1 Dataset

The dataset used for the proposed research expanded the babyPose-v1 by adding 11 depth videos, for a total of 27 depth videos.

For each video 1000 frames were annotated and were randomly split into training and testing data. 750 frames were used for training and validation purposes and the remaining 250 frames were used to test the network. Unlike the previous 2D and

Table 2.4: Ablation study for the DeA detection network. The table shows also the number of trainable parameters fot each of the architecture.

	Architecture	Parameters	Ablated component	
			Enc2-Dec2	**Enc3-Dec1**
	Proposed DeA detection	20881532	flowing pathway	long-skip
1)	2D detection CNN	15529060	long-skip	long-skip
2)	DeA detection-1	37010300	long-skip	flowing pathway
3)	DeA detection-2	42391164	flowing pathway	flowing pathway

3D pipeline for each infant, supported by the SINC clinical partners, the training and testing frames were divided so as to maximize the differences between the two sets, particularly in terms of poses and occlusions. The resulting dataset accounts 20250 frames to train and validate and 6750 frames to test the architectures.

2.4.2.2 Ablation study

To prove the potentiality of the flowing pathway the performance of the DeA detection network was compared against that of the 2D detection CNN. This network shared the same architectural design of the DeA detection network without including the DeA pathway.

A further ablation study was conducted by varying the position of the flowing pathway between Enc3-Dec1 while keeping the long-skip connection between Enc2-Dec2. Moreover, inspired by [1], two flowing pathways were implemented both between Enc2-Dec2 and Enc3-Dec1. Table 2.4 summarizes the ablation studies conducted for the DeA detection network.

To prove the effectiveness of the densely connected atrous-based pathway ablation studies for the DeA regression network were lead and its performance was compared against those of its closest variant (i.e., the 2D regression CNN).

2.4.2.3 Investigation on DeA detection efficiency

Most of the research on developing CNNs to solve the task of detection, focuses on improving estimation accuracy with few consideration the model efficiency. Developing increasingly sustainable models is crucial to ensure on-the-edge computing especially in scenarios where computational resources may not be available (e.g., in a domestic environment). Following the considerations in [1] this work sought for a balance between network complexity and improved performance. Thus, inspired by [68, 69], variants of the DeA detection network based on asymmetric convolutions were implemented. Asymmetric convolutions reduces the architectural complexity while attaining almost unaltered performance. This is achieved by splitting, for example, the traditional 3x3 convolution into two cascaded asymmetric convolutions of kernel sizes equal to 3x1 and 1x3, respectively [68]. As showed in Table 2.5, firstly

Table 2.5: Comparisons for optimizing model efficiency. The table reports the names of the newly stated architectures, their trainable parameters and the architectural component where we applied asymmetric convolutions. The DeA detection network was compared against its asymmetric variations. Asy-DeA-Enc detection implements the original decoder (Dec) and DeA pathways while keeping the encoder (Enc) asymmetric. Asy-DeA-body implements the original DeA pathways while making asymmetric both the Enc and the Dec. The Asy-DeA detection represents the asymmetric version of the DeA detection.

	Architecture	Parameters	Asymmmetric component		
			Enc	**Dec**	**DeA pathways**
	Proposed DeA detection	20881532			
1)	Asy-DeA-Enc detection	18806460	✓		
2)	Asy-DeA-Body detection	18285180	✓	✓	
3)	Asy-DeA detection	16522650	✓	✓	✓

the asymmetric version of the DeA detection network (i.e., Asy-DeA detection) was implemented, this network had the asymmetric convolutions both in the Enc and Dec and in the DeA pathways. The second architecture, named Asy-DeA-body detection, leveraged asymmetric convolutions both on the Enc and Dec while keeping traditional convolutions in the DeA pathways. The last tested architecture implemented asymmetric convolutions in the Enc only, thus we called it Asy-DeA-Enc detection.

2.4.2.4 Training settings and performance metrics

All the tested architectures follow the training settings presented in Sec. 2.3.2.2. Frames were resized to 128x96 pixels in order to smooth noise and reduce both training time and memory requirements. Mean intensity was removed from each frame. The detection-network family used Adam as optimizer and the per-pixel binary cross-entropy (i.e., L_{CE}) as loss function. While the tested regression CNNs were trained with the stochastic gradient descent as optimizer with the mean squared error (i.e., L_{MSE}) as loss function. Unlike the previous 3D pipeline, these losses have been readjusted for the single-frame analysis task.

Similarly for the training settings, also the performance metrics used to evaluate the approaches were the same of the 3D pipeline described in Sec. 2.3.2.2. Namely, the *DSC* and the *Rec*, for the detection networks, and the *RMSD* for pose performance estimation.

To accomplish the experiments the same hardware presented in Sec. 2.3.2.2 was used (i.e., Intel® Xeon® Silver 4214 CPU @ 2.20GHz with 230 GB of RAM and a NVIDIA® RTX 2080 8 GB RAM).

Table 2.6: Joint-connection- and joint- detection performance in terms of median recall (*Rec*) and Dice Similarity Coefficient (*DSC*).

Architectures	Median *Rec* for joint-connections	Median *DSC* for joints	Median *DSC* for joint-connection
Proposed DeA detection	0.89	**0.90**	0.89
DeA detection-1	**0.89**	0.89	**0.90**
DeA detection-2	0.88	0.89	0.89
2D detection CNN	0.87	0.89	0.89

2.4.3 Results

The results for the ablation studies are reported in the boxplots showed in Fig. 2.17 and in Table 2.6. The boxplot shows the recall achieved by the tested architectures when detecting the joints. The three DeA architectures achieved similar results, the highest median *Rec* for joints prediction was achieved by the DeA detection network (median *Rec*=0.89) while the lowest by the DeA detection-2 (median *Rec*=0.88). The 2D detection CNN attained the lowest performance with respect to the 3 DeA-based architectures (median *Rec*=0.86). This performance trend is also reflected in the results shown in the Table 2.6 which reported the median *Rec* for the joint-connections and the median *DSC* for the joints and joint-connections, respectively.

Further investigation to find an effective and efficient model was conducted on the DeA detection CNN which, among its 3 ablated versions, combined accuracy in prediction and fewer parameters (DeA detection $\sim$20M parameters, DeA detection-1 $\sim$ 37M parameters, DeA detection-2 $\sim$ 42M parameters). This study was conducted to seek for reducing network computational complexity while keeping its performance almost unchanged. Fig. 2.18 showed the boxplots of the *Rec* achieved by the architectures. All the architectures achieved similar results with respect to the original DeA detection network. The Asy-DeA-Enc detection which implemented asymmetric convolutions in the Enc only, achieved the highest results with a median *Rec*= 0.894.

In support of quantitative results, samples of predictions on challenging frames for the DeA detection network (second column) and the 2D detection CNN (third column) are shown in Figure 2.19. The white arrows highlighted detection errors committed by the 2D detection CNN while the white cross means that no predictions occurred.

The performance comparison in terms of *RMSD* of the DeA regression network and the 2D regression network are shown Fig. 2.20. The DeA regression network achieved the highest performance with a median *RMSD*=10.789 pixels among the four limbs while the 2D regression CNN obtained a median *RMSD*=11.269 pixels. Figure 2.21 showed the results achieved by the DeA regression network on frames with different occlusions (i.e., self or external occlusions) or mispositioning of the acquisition set-up.

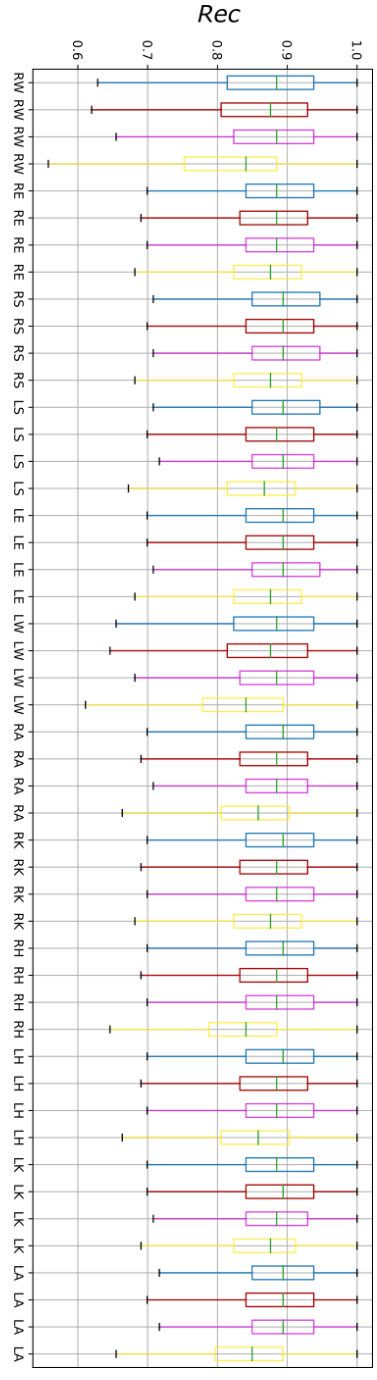


Figure 2.17: Boxplot of the Recall (*Rec*) for joints prediction. These results were achieved by the network tested for the ablation studies summarized in Table 2.4. Blue: results achieved by the proposed DeA detection network; Red, pink and yellow: results achieved by the DeA detection-1, DeA detection-2 and the 2D detection CNN, respectively. In green the median value was reported.

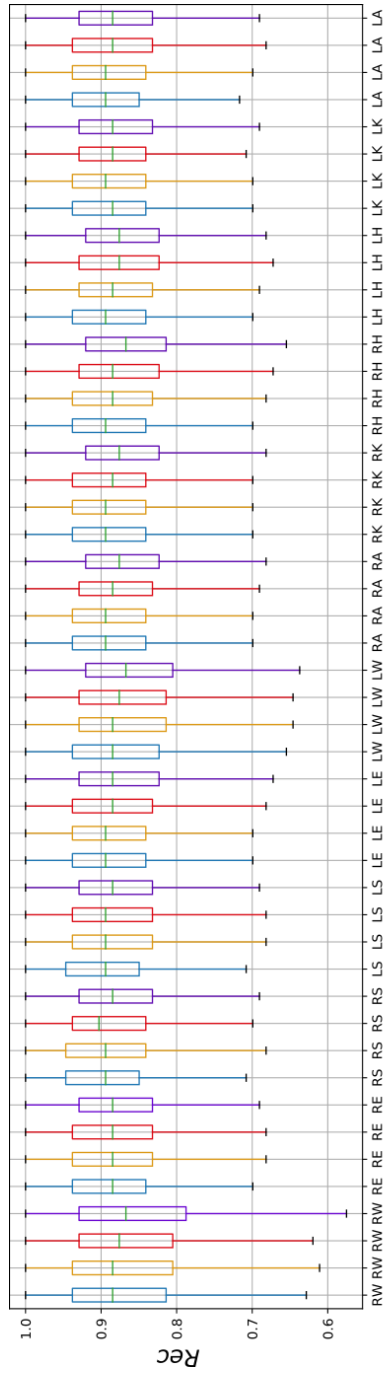


Figure 2.18: Boxplot of the recall (*Rec*) for joints prediction. These results were achieved by the networks tested for further investigation on DeA-detection efficiency (as reported in Table 2.5). Blue: results achieved by the proposed DeA detection network; orange, red and purple: results achieved by the Asy-DeA-Enc detection, Asy-Dea-Body detection, Asy-DeA detection, respectively. In green the median value was reported.

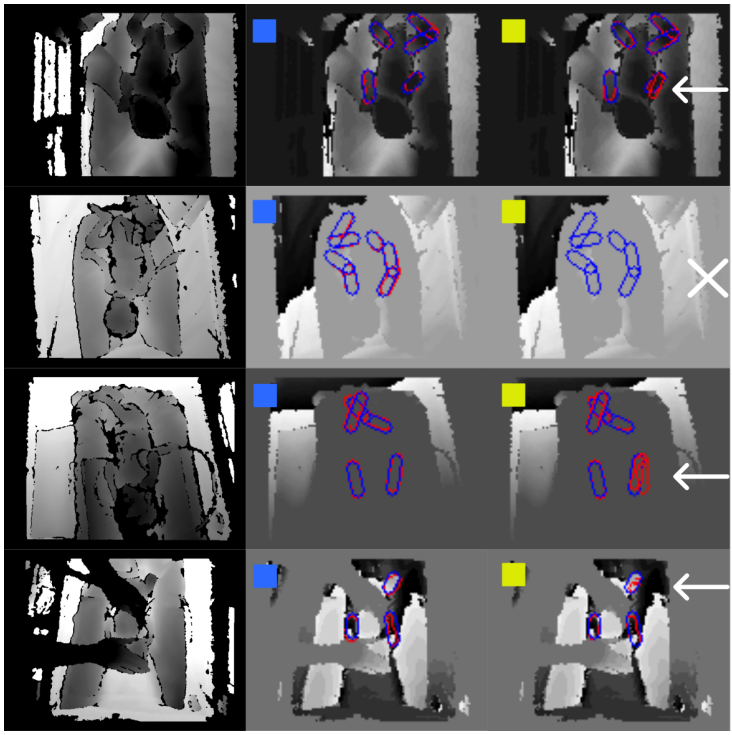


Figure 2.19: Samples of qualitative results in challenging frames for the DeA detection network (second column, images marked with a blue square) and the 2D detection CNN (third column, images marked with a yellow square). The first column shows the original images while the second and the third column represent the predictions of the network (red) and the corresponding ground truth (blue) superimposed to the preprocessed images. The white arrows in the third column highlight the prediction errors committed by the 2D detection CNN while the cross stands for no predictions.

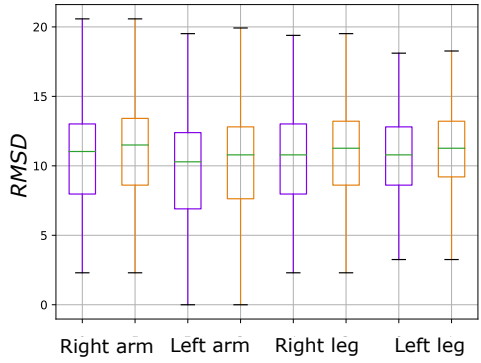


Figure 2.20: Boxplot of the root mean square distance error *RMSD* calculated for each of the four limbs. The violet boxplots show the results achieved by the the DeA regression network while the orange the results of the 2D regression CNN. In green the median value was reported.

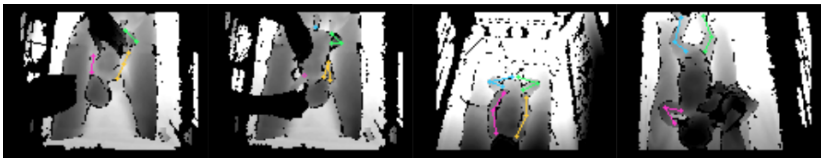


Figure 2.21: Samples of qualitative results achieved by the DeA-regression network when estimating the limb-pose. Limb-pose was superimposed to the original depth frame.

2.4.4 Discussion

The proposed DeA detection CNN achieved encouraging results as shown in Fig. 2.17 with a median *Rec* for the joints of 0.89, overcoming the 2D detection network which achieved a median *Rec* of 0.86.

Accompanying quantitative results in validating the research hypothesis, qualitative results on challenging frames are showed too (Fig. 2.19). These qualitative results visually show the predictions of both network architectures superimposed on challenging frames (i.e., frames with occlusions or homogeneous pixel intensity, etc). As visible from the figure, the DeA detection network better delineates the infants' limbs in challenging frames. This suggests that introducing the flowing pathway enables the Dec path to be enriched with all the information content lost during the information-shrinkage operations carried out by the Enc path.

To look for the optimal network configuration, the DeA detection network underwent different architectural variations (Table 2.4). The results showed in Table 2.6 and Fig. 2.17 suggested that all the 3 DeA detection-based architectures performed similarly and achieved increased performance with respect to the 2D detection CNN. The lowest performance when compared the 3 DeA detection-based architectures where achieved by the DeA detection-2 which was inspired by [1]. This network, which implemented two flowing pathways, turned out to be complex enough to memorize the training data and consequently loosing its effective capacity [70].

As claimed by [1], when dealing with CNNs which must be deployed in scenarios where computational resources may not be guaranteed, considerations relevant to model efficiency are crucial. Guided by this hypothesis further studies were conducted on DeA detection model efficiency which had $\sim$17M fewer parameters with respect to the DeA detection-1. These studies are aimed at investigating possible architectural variations that would make the architecture lightweight while attaining unaltered its performance. To this goal asymmetric convolutions [71] were implemented in the 3 main network parts (i.e., Enc, Dec, DeA pathways). These newly stated architectures are reported in Table 2.5. As showed by the boxplot in Fig. 2.18 the Asy-DeA-Enc detection achieved the same median *Rec* (=0.89) of the DeA detection while the lowest performance where achieved by the Asy-DeA-Body and Asy-DeA detection with median *Rec* for joints equal to 0.88 and 0.88, respectively. As pointed out by the authors in [71], this decrease in performance was induced by the increase in network depth resulting from the implementation of asymmetric convolutions which replace the traditional convolution (e.g., kernel size= 3x3) in two consecutive convolutions (e.g., kernel size= 3x1 and kernel size= 1x3). Hence, when training a network from scratch, increasing its depth may pose issue related to the vanishing gradient [71]. Consequently this outcome did not deal with the Asy-DeA-Enc as the network implemented asymmetric convolutions in the Enc part only, slightly increasing the DeA detection depth.

Concerning the DeA regression network, its performance was compared with the

ones of the 2D regression CNN (which is the closest network architecture to the one proposed). The DeA regression network achieved the highest performance in terms of *RMSD*. The quantitative results coupled with qualitative ones (Fig. 2.21) suggested that guiding DeA regression with information collected from DeA detection and still implementing DeA layers in the former architecture may guide the network in a more precise joint and joint-connection localization particularly in images with inherent (e.g., noisy image portion, joint self-occlusions) or extrinsic (e.g., external occlusions, mispositioning of the acquisition set-up) challenges.

A limitation of the work may be seen in the test set and training set consisting of different portions of the same video. This issue stems from the fact that the data used to train and validate the approach are characterized by an high variability, mainly induced by: (i) the differences between preterm infants in terms of gestational age, weight and height, (ii) the different clinical conditions of the infants, (iii) the presence of external occlusions while collecting data (e.g., sheets, pillows, splints, therapy equipment or the hands of the operators and parents...). Publicly releasing the babyPose-v2 will certainly encourage other researchers to take on the inherent challenges of this research and to propose increasingly reliable and affordable monitoring systems.

Monitoring preterm infants' GMs, directly in NICUs, has a strong predictive value for early diagnosing the presence neurodevelopmental disorders. As highlighted [47], pursuing research in preterm infants' pose estimation is fundamental to ensure reliable and accessible models which guarantee continuity of care both in NICU and even after the hospitalization. This work therefore takes conscience of the necessities in such a delicate field as that of preterm infants' care, and proposes a lighter version of the pipeline based on 3D convolutions while designing architectural variations able at improving the performance of the 2D-convolutions-based pipeline. The topics covered in this research may open new research scenarios aimed at developing embedded monitoring solutions for on-the-edge computation. This would both break down cost constraints in terms of computation and money, and ensure that such increasingly advanced monitoring solutions can be deployed anywhere, without barriers.

2.5 A sustainable deep learning approach for preterm infants limbs' detection

Innovative technologies open up new possibilities that require the scientific community to completely rethink the ways to approach research in e-health [72]. Promote and actively support specific actions aimed at raising awareness in sustainable health issues constitutes a crucial research paradigm [5]. Monitoring applications as those proposed in the previous sections (i.e., Sec. 2.3 and Sec. 2.4) represent a clinical and therapeutic breakthrough towards a more personalised medicine. However this innovation must be accompanied by more responsibility in designing and developing accessible DL-based applications.

Indeed, both the presented approaches proved to be effective in the task of preterm infants' pose estimation but considerations about the efficiency of such models were barely advanced with the DeA-based pipeline. As claimed in [73], designing computationally intensive DL models (i.e., with dozens of millions of trainable parameters) requires demanding computational, memory and energy resources, limiting the applicability of such monitoring solutions to computationally resource-intensive scenarios.

In this research, responsible innovation in the medical field translates into the efficiency enhancement of available resources. The innovative elements of the research are summarized hereafter:

1. The design of the TwinEDA, a CNN conceived for being sustainable. TwinEDA is intended to perform similarly in terms of efficacy with respect to the 2D detection network, while being more efficient in terms of computational and memory resources requested;

2. Deployment on-the-edge of TwinEDA via a SBC device. This allowed to assess the viability of using our DL-based monitoring system in clinical settings with limited resources (in terms of computational power and memory); The chosen device is the NVIDIA® Jetson Nano, a hardware platform oriented to DL that aims to democratize AI [74].

2.5.1 Efficiency in convolutional neural networks as a mean to improve sustainability

In today's DL-centric research paradigm, every advance is primarily achieved via: increasingly large datasets to guarantee better generalisation power to the networks and increasingly complex and inefficient models at the operational level (i.e., models with an increasing size and number of trainable parameters) [75]. These development choices, however, have an unavoidable impact on the energy and economic sustainability of the solutions. In fact, increasingly complex models require longer

information-processing times, with a decisive impact on the energy needed to support the computation [76]. As a consequence, the computational resources required to deploy these solutions must be more and more performing (and expensive), thus colliding with the responsibility of those who conduct research in the clinical field: to create systems that can be distributed without borders and without limits imposed by the absence of economic and computational resources [5].

Taking responsibility for making CNN models increasingly efficient (and, therefore, sustainable) and possibly embed such models in inexpensive and environmentally friendly computing devices, are crucial challenges to ensure that everyone enjoy the best possible medical treatment.

The need to push the research towards the development of more sustainable DL models has been echoed by other researchers in closer fields [73]. On-the-edge DL applications [77, 78, 79] became pervasive in many real-world scenarios (e.g., video-surveillance systems or self-driving cars) which required efficient applications for objects detection and classification or people and people's behaviour identification. Following the paradigm of on-the-edge DL applications, the authors in [2] proposed EDANet, a CNN for real-time semantic segmentation. EDANet has some architectural peculiarities that guarantee efficiency while ensuring quite accurate performance. EDA modules bundled 3 main architectural choices aimed at lowering network complexity:

- **Asymmetric convolutions** [69] break the 2D convolution (kernel size = nxn) into two cascaded 1D convolutions (kernel sizes = 1xn, nx1), saving significantly the number of trainable parameters while keeping almost unaltered CNN performance.

- **Atrous convolutions** [66] as shown in Sec. 2.4, these convolutions space out kernel values by d zeros, where d is the dilation rate. This design enlarges the receptive field of the kernel without increasing the number of trainable parameters.

- **Densely connected layers** [65] allow the deepest layers of the network to receive the features of all the shallower layers. Thus subsequent layers are responsible for learning few new features.

Unlike Unet-shaped networks (such as 3D, 2D and DeA detection CNNs), EDANet replaces the decoder with a bilinear interpolation block for upsampling features. Bilinear interpolation is not a data-driven operation, considerably decreasing the number of trainable parameters.

Probing this line of research, the proposed work combines the architectural choices of EDANet [2] and the 2D detection CNN. EDANet, although computationally sustainable, performs too poorly to be translated in the clinical practice, while the 2D detection CNN reaches accurate performance without caring for the efficiency. Coupling

the strengths of both CNNs allows to define TwinEDA network that simultaneously ensures efficiency and efficacy.

2.5.2 Methods

2.5.2.1 TwinEDA convolutional neural network

Figure 2.22 shows the TwinEDA network along with the two CNNs (i.e., EDANet and 2D detection CNN) chosen as baseline references for designing an architecture that is both effective and efficient. TwinEDA network combines both the architectural peculiarities of the 2D detection CNN, which has been designed for effective preterm infants' joint and joint-connection detection, and the EDANet, thought for being sustainable (in terms of trainable parameters and memory requirements) and performing real-time semantic segmentation.

As in EDANet, TwinEDA, downsamples the input image via two bi-branch down sampling blocks which concurrently exploit 3x3 strided convolutions (with stride (s)=2) and maxpooling layers to lessen the computational cost and gather compact contextual information. The output of the downsampling blocks enters the first two parallel EDA modules which smoothly process joint and joint-connection information. EDA modules consist of six and five densely connected sub-blocks, respectively. Each sub-block stacks a 1x1 convolution (to reduce the number of channels), an asymmetric 3x3 convolution and a atrous 3x3 asymmetric convolution. Asymmetric atrous convolution progressively increases the d factor (d=1,2,4,8), this widens the filter field of view allowing the retrieval of multiple scale information of the image while keeping low the number of learnable parameters. In each EDA module, the data flows in a dense fashion to strengthen CNN ability in expressing features while smoothing information flow [65]. Outputs from each EDA module are combined prior entering two parallel downsampling blocks which capture coarse contextual information for joint and joint-connections prior entering further two parallel EDA modules. TwinEDA shares the same paradigm of the 2D detection network, by involving chain of bi-branch blocks reuniting in a single convolutional layer and newly forked. This design ensures the parallel processing of joints and joint-connections without losing the intrinsic continuity of these anatomical structures within the human body [58].

To restore the information up to the output layer, the upsampling part of the TwinEDA shares the same bi-branch structure of the downsampling one. It combines a single layer of 1x1 convolution and then two parallel layers, one of which implements a 3x3 up-convolution and the other one a bilinear interpolation operation. This choice relies upon the hypothesis that bilinear interpolation saves computation but rises issues relevant to inaccurate joint and joint-connection detection [2]. On the other hand, the upconvolution blocks proposed in the 2D detection CNN are gradually trained to precisely outline joint and joint-connections localization but are computationally burdensome operations. Combining demand-driven and data-driven operations solves the

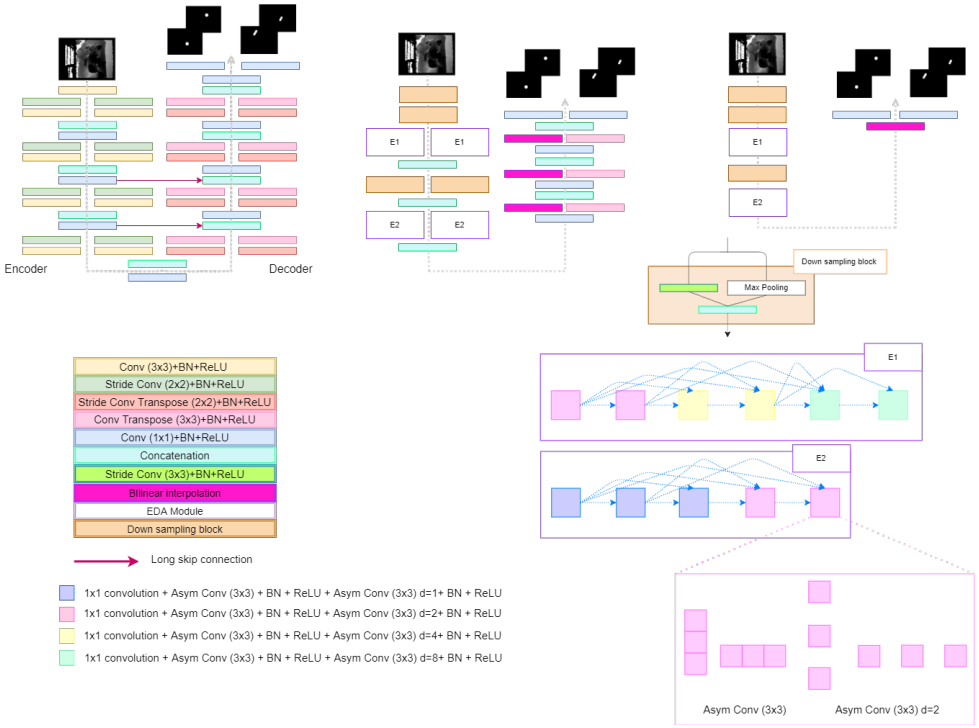


Figure 2.22: Structure of detection architectures: 2D detection convolutional neural network (left), the TwinEDA (center) and the EDANet in [2]. TwinEDA combines the architectural choices of the 2D detection network (e.g., bi-branch architecture to process joints and connections in parallel) and the EDANet network (e.g., EDA modules) [2] to ensure an effective and efficient architecture. The color-caption is shown in the bottom left corner of the figure. The dashed grey arrows highlight the direction of information flow along the networks.

issues posed by the individual implementation of the two operations.

As for the 2D, 3D and DeA detection networks, to train the TwinEDA the 20 affinity binary maps (i.e., 12 for joints and 8 for joint-connections) were prepared.

2.5.2.2 Deployment

The TwinEDA network has been deployed on the NVIDIA® Jetson Nano, a portable and cost-effective hardware. The Jetson Nano used for the deployment couples 4 GB RAM, 4-cores ARM A57 CPU with an on-board GPU with 128 CUDA cores based on a Maxwell microarchitecture design.

To evaluate the performance of the network over the NVIDIA® Jetson Nano in terms of prediction time for a single-depth frame, a two-step process was accomplished, including: (i) conversion of the model from the original format (i.e., Tensorflow) to the *.onnx* [5] format and (ii) a subsequent conversion to the serialized *TensorRT*™ [6] engine format. *TensorRT*™ generates an optimized version of the TwinEDA model to ensure efficient performance when deployed on the Jetson device.

2.5.3 Experimental Protocol

2.5.3.1 Dataset, training settings, performance metrics

For the experiments the babyPose-v2 dataset was used, keeping the same division into training, testing and validation set proposed in Sec. 2.4.2.3 (i.e., 750 frames per infant to train and validate the CNN and 250 frames to test it).

As in Sec. 2.4.2.3 and Sec. 2.3.2.2, the optimal combination of loss, learning-rate scheduling and optimizer, was found after a grid-search analysis. As a result of this analysis, the CNN was trained for 100 epochs with Adam as optimizer and the per-pixel binary cross entropy (i.e., L_{CE}) as loss function. Concerning the performance metrics, besides *DSC* and *Rec*, inspired by [73], to quantitatively evaluate the efficiency of each model, the number of trainable parameters, the inference speed and the model memory requirements were computed. The inference speed was assessed both on the same hardware in Sec. 2.3.2.2 and Sec.2.4.2.4 and when deploying the TwinEDA on the NVIDIA® Jetson Nano.

2.5.3.2 Investigation on sustainability and ablation studies

As claimed in a recent contribution [80], the entire process of developing an AI application, from the idea, through the design of the architectures to the deployment phase of the models, should embrace the paradigm of sustainability. This research hypothesis is even more crucial if applied in such a delicate field as medicine, where patients need to be guaranteed high-quality care. To this goal both the efficiency and efficacy

[5] https://onnx.ai/
[6] https://developer.nvidia.com/tensorrt

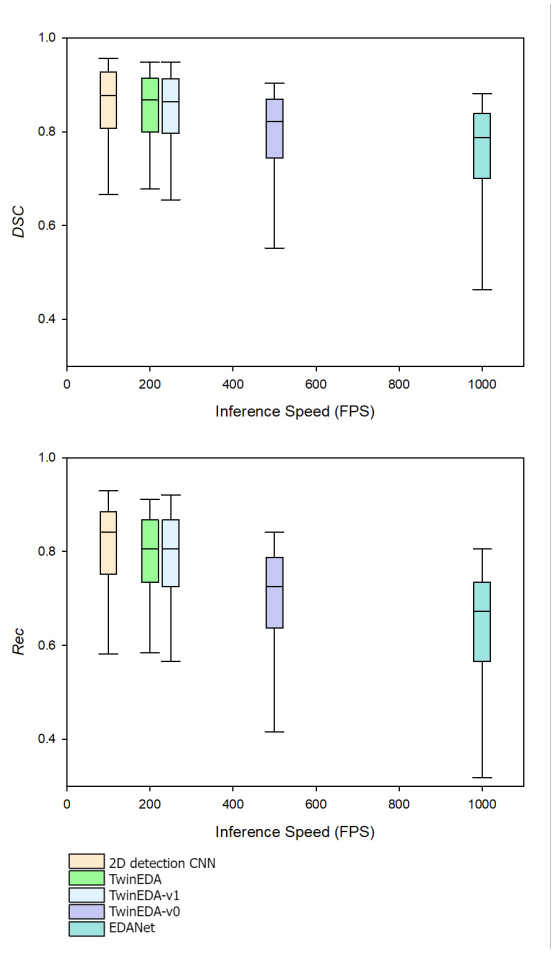


Figure 2.23: Boxplots for quantitatively evaluate the efficiency and efficacy of the network when detecting the joints. The x-axis shows the network inference rate in terms of frames per second (FPS). The y-axis shows the performance in terms of Dice similarity coefficient (*DSC*) (top) and Recall (*Rec*) (bottom) for mean-joint-detection. The different colors of the bloxplots represent the different architectures while the black line depicts the median values, the caption is shown at the bottom of the image.

Table 2.7: Sustainable investigation: the performance of the TwinEDA network was compared with that of the two networks chosen as baselines for its design (i.e., EDANet [2] and the 2D detection CNN), TwinEDA-v0 and TwinEDA-v1. The table reported for each architecture the number of trainable parameters.

Architecture	Parameters
TwinEDA	2.66M
EDANet	620K
TwinEDA-v0	1.61M
TwinEDA-v1	3.29 M
2D detection CNN	15.5M

of the TwinEDA network were evaluated and compared against those of closer architectures. Thus, the sustainable investigation was conducted on three architectures: the TwinEDA, EDANet and the 2D detection CNN.

To prove the effectiveness of coupling demand and data-driven operations, as an ablation study, two variants (among all the tests performed to find the optimal architecture) of the TwinEDA were tested. TwinEDA-v0 and TwinEDA-v1 share the same downsampling path of the TwinEDA and differ from each other for the upsampling path. Hence, to upsample data, TwinEDA-v0 implements bilinear interpolation while TwinEDA-v1 exploits subsequent up-convolutional blocks (with kernel size= 3x3) prior entering the output layer. The number of parameters for each of the CNN is shown in Tab. 2.7.

2.5.4 Results

To assess the performance of the architectures in terms of efficacy we computed the median values in terms of *DSC* and *Rec* for joints and joint-connections and reported the results in table 2.8. TwinEDA achieves closer results to the 2D detection CNN, with the same median values of *DSC* for joint-connections detection equal to 0.89. The boxplots in figure 2.23 both show performance for joints detection in terms of *DSC* and *Rec* and highlight the inference speed (in terms of frames per second) for each of the architectures.

These quantitative results demonstrate how Twin architectures (particularly TwinEDA and TwinEDA-v1) achieve similar results to the 2D detection CNN while the lowest performance was achieved by EDANet and TwinEDA-v0. Such a trend is also reflected in the boxplots (figure 2.23) which combine quantitative performance with information on the efficiency of CNNs in terms of inference time.

Regarding efficiency considerations, as shown in the boxplots and reported in Table 2.9, the TwinEDA network reduces the prediction time for a single frame by an order of magnitude with respect to the 2D detection network (i.e., prediction latency of the 2D detection CNN=0.01 s and prediction latency of the TwinEDA= 0.005 s).

Table 2.8: Quantitative results for joints' and connections' detection in terms of median Dice similarity coefficient (*DSC*) and Recall (*Rec*). Interquartile ranges for each measure are shown in parentheses.

Architecture	Median *DSC*		Median *Rec*	
	joint-connections	joints	joint-connections	joints
TwinEDA	0.89 (0.08)	0.88 (0.09)	0.86 (0.11)	0.83 (0.10)
EDANet	0.79 (0.11)	0.81 (0.09)	0.69 (0.16)	0.70 (0.12)
TwinEDA-v0	0.80 (0.11)	0.83 (0.09)	0.72 (0.16)	0.73 (0.12)
TwinEDA-v1	0.88 (0.08)	0.87 (0.09)	0.85 (0.12)	0.81 (0.10)
2D detection CNN	0.89 (0.08)	0.89 (0.09)	0.87 (0.11)	0.86 (0.09)

Table 2.9: Model efficiency assessments in terms of inference speed (i.e., the time to predict a single depth frame) for all the 5 tested architectures and memory requirements occupied by each model.

Architecture	Inference speed [s]	Model size [MB]
TwinEDA	0.005	46.2
EDANet	0.001	8.2
TwinEDA-v0	0.002	21.5
TwinEDA-v1	0.004	60.3
2D detection CNN	0.010	186.7

Moreover, TwinEDA model has a size of 46.2 MB which, compared to the 2D detection CNN (model size=186.7 MB), is considerably smaller.

The TwinEDA was further deployed on the NVIDIA® Jetson Nano. The network takes 0.05 s to process a single frame, paving the way to making these monitoring devices usable in clinical practice.

2.5.5 Discussion

This research was aimed at designing a DL based monitoring system able to accurately monitor the infants' limbs-movements with special care paid to the sustainability of such systems in terms of computational and, consequently, economic resources employed. With this aim, the TwinEDA network for preterm infants' movement monitoring was proposed to ensure effective predictions and computational efficiency.

As shown in Table 2.8 and Fig. 2.23, the proposed TwinEDA performed similarly with respect to the detection network in [81] with the same median *DSC* and an absolute difference of one percentage point between the two median *Rec* when detecting joint-connections. Combining quantitative results with efficiency performance (Figure 2.23 and Table 2.9) it emerges that the 2D detection CNN is dramatically slower in predicting a frame (0.010 s) than the TwinEDA network (0.005 s) and takes up one more order of magnitude of memory space (TwinEDA network is 140.5 MB smaller in size and has 12.5 M parameters less). These results lead to two considerations: (i) the complex and deep nature of the 2D detection network makes it inefficient and computationally prohibitive and (ii) the bi-branch structure that guarantees parallel processing of joints and connections allows to precisely trace joints and joint-connection in space.

EDANet represents the boundary for temporal efficiency. The TwinEDA CNN, compared to EDANet, is slower when predicting a single frame (it takes 0.004 s longer), has a model size and number of trainable parameters which, although contained, are slightly higher (TwinEDA network is 38 MB larger in size and has more than 2 million parameters more). As far as comparing the performance achieved by the networks in finding the position of joints and connections, the TwinEDA network

achieved significantly improving results (median *DSC* and *Rec* for joint-connections detection of 10 and 17 percentage points higher, respectively). This improvement in performance is driven by the combination of data-driven and demand-driven operations. Indeed, as previously pointed out [2], information upsampling performed via operations of demand-driven nature only, reduces the amount of computation but does not properly reconstruct the information lost during the shrinkage operations in the downsampling phase.

Further comparisons were conducted to validate the TwinEDA architecture as a trade-off between quality in results and low computational demand. To this goal the performance of the TwinEDA network against the ones of its twins (namely TwinEDA-v0 and TwinEDA-v1) was assessed. Of the 3 Twin-architectures, TwinEDA-v0, is the most efficient with a model size equal to 21.5 MB, a number of trainable parameters equal to 1.61 M and a latency time to predict a single frame of 0.002 s. However, as depicted in Fig. 2.23 and reported in Table 2.8, the network, compared to TwinEDA, is ineffective when detecting the positions of the joint and joint-connections (median *DSC* and median *Rec* for joint-connections detection equal to 0.80 and 0.72, respectively). This further proves that keeping bilinear interpolation in the upsampling path may limit CNN capabilities in detecting infants' limbs. To solve the issue of the upsampling path, the performance of TwinEDA-v1 (which implements subsequent mono-branch convolution operations) was tested. Introducing data-driven operations enable the TwinEDA-v1 to perform similarly with respect to TwinEDA (median *DSC* and *Rec* one percentage point lower) and better with respect to the TwinEDA-v0 (median *DSC* and *Rec* 8 and 13 percentage points higher, respectively). However, compared to the TwinEDA, which parallels data- and demand-driven operations, TwinEDA-v1, is more computationally demanding (model size= 60.3 MB and number of trainable parameters= 3.29 M). The slightly drop in latency time, with respect to the TwinEDA, is due to a mild decrease in network depth.

Promoting equity of access to high-quality care is of primary importance, so to reinforce the need of making innovative technologies accessible to all, the TwinEDA network was deployed on the NVIDIA® Jetson Nano. This device for on-the-edge computing was designed to make AI applications computationally and cost-wise sustainable. This test showed that the designed TwinEDA can be optimized and deployed on this computing device and achieves acceptable performance in terms of single-frame prediction time (i.e., inference speed= 0.05 s, 20 FPS, real-time working threshold=0.08s, 30 FPS).

This research reframes some concepts of the development of DL models in a more sustainable and modern way. It tries to raise the need to propose clinical applications that guarantee effectiveness along-side with efficiency. Indeed, as highlighted in [5] designing reliable and affordable digital technologies will be a transformative force in the healthcare sector and this work is among the first to underline the importance of such a paradigm shift.

2.6 Conclusion and future perspective

Monitoring preterm infants' limb-movement, directly in NICUs, has a strong predictive value for early diagnosing the presence neurodevelopmental disorders. Despite its relevance, monitoring spontaneous motility is mainly performed visually by trained clinicians with the drawbacks of being subjective, time-consuming and qualitative. To solve issues due to the visual assessments, authors in literature proposed to monitor movements via wearable sensors. However, adding extra burden to a fragile and extremely worn out body may cause additional discomfort, pain, itching. Moreover wearable sensors may hinder infants' spontaneous motility and may suffer from artifacts caused by the movement of healthcare operators and parents interacting with the child. Vision-based systems may represent a viable alternative to wearable sensors, allowing the monitoring of infant's spontaneous motility without being in contact with infant's skin and leaving operators and parents free to move around the crib. In the literature, vision-based approaches applied to the monitoring of preterm infants are mostly based on the analysis of RGB (e.g., [44]) posing issues relevant to privacy.

With the view to overcome state-of-art limitations this chapter presents DL-based pipelines which analyse depth clips or images collected during the actual clinical practice (Sec. 2.2). The chapter collects the story of a three-year journey and bears witness to a maturation of awareness. The first DL pipeline (Sec. 2.3) exploited 3D convolutions to estimate limb-pose from depth streams. The 3D pipeline, compared with its akin which exploited 2D convolutions for single-depth frames analysis, proved to be effective in the task of pose estimation. However, the computational complexity introduced by 3D convolutions is prohibitive and makes the monitoring system unsuitable for being translated in the actual clinical practice. The initial need slightly changes and the awareness of paying more care of the energy consumption of DL-based monitoring systems matures.

The 2D pipeline (Sec. 2.3), which is naturally more efficient than the 3D one, was reviewed and architectural variations were implemented to improve its performance (Sec. 2.4). Architectural modification (e.g., asymmetric convolutions, atrous convolutions...) to lower the computation effort of the 3D pipeline while improving the generalization power of the 2D one were investigated. This newly DeA-pipeline got improved quantitative performance with respect to the 2D pipeline while being more efficient with respect to the 3D one. Proving, for the first time, that starting to think about the efficiency of neural networks is critical to large-scale deploying such innovative monitoring solutions.

The first sustainable DL architecture for preterm infants' limb-detection (Sec. 2.5) was proposed to further lower the computational cost and memory consumption of the 2D pipeline. The designed TwinEDA follows two main reflections: (i) health is of general interest and any innovation should be distributed worldwide, without borders, thus extending advanced monitoring systems even in scenarios where computational

(and, therefore, economic) resources are not guaranteed became more and more crucial; (ii) as researchers we should rethink the way we relate to neural-network design, which must be increasingly aimed at proposing sustainable solutions to limit energy consumption. As a result, TwinEDA guarantees both the efficiency performance of the 2D pipeline but largely decreased the computational demand as to be deployed on a lightweight and cost-effective NVIDIA® Jetson Nano.

The approaches presented in this chapter lay the foundations for a new paradigm of infants' monitoring in NICUs, showing promising results. Further research is needed to realize an integrated systems that can be optimal and widely usable both in clinical practice and at home after the hospitalization. This work will pave the way for the development of computer-assisted diagnosis tools to better analysing GMs and timely recognizing signs of the presence of neurobehavioral pathologies as ASD.

Chapter 3

Automatic assessment of the autistic child's autonomy in daily actions

The child with ASD is characterised by a different development of the central nervous system than what is considered normal. This is partly based on genetics and on environmental factors. The environment causes in these children the inability to find the right stimuli for their diversity. This increases, in the children, the difficulties of socialisation and ends up in creating a bubble in which ASD children find themselves enclosed. With no possibility of communicating outside this bubble. This generates suffering above all in the children and then in their families. The sooner this neurodiversity, which differs from child to child, is understood the more the child grows and develops his/her potential.

The ABA therapy is currently the most effective in the treatment of children with ASD. At the basis are artificial stimuli and reinforcement, to teach the child to behave and interact with the surrounding environment or to reduce self-injurious and repetitive behaviours. By treating each symptom, each error in the metabolic system with appropriate tests, a gradual change and improvement in children's health and behaviour verifies.

The research here presented falls within the "COMEACASA" project which deals with the proposal of innovative intervention methodologies for children with autism. It stems from a dialogue with ABA operators from a centre specialised in ABA therapy located in Macerata (Italy). Indeed, ABA therapy is not merely applicative but involves extensive monitoring of the child and his/her behaviour in relation to environmental stimuli. To understand the child's specific neurodiversity, the operators observe the child while he/she is carrying out specific tasks and design children-specific programs. To provide ABA operators with a non-intrusive support ally, the same camera used to monitor preterm infants in cribs was mounted over the bathroom sink. The video recordings were used to quantify the children's autonomy in the act of washing their hands, which, of all actions, as we have learned during the COVID-19 pandemic, is crucial to the safety of the children and those around them.

3.1 Background and motivation

Figure 3.1: The acquisition set-up (white square, left), placed in the bathroom to record the sink (pink square, right), consisted of an Astra Mini S-Orbbec® RGB-D camera and a minipc Intel® NUC core *i5*.

Autism is a severe permanent neurodevelopmental disorder with long-term and pervasive effects. It affects 1 in 59 children, worldwide[1]. Impaired communication skills, inability to socially interact, absence of emotional reciprocity, limited interests and repetitive behaviors are among the major adverse implications of ASD [82].

The treatment of children suffering from ASD mostly relies upon the ABA technique. The ABA aims at modifying the child behaviour to make it functional to the tasks of everyday life (e.g., nutrition, personal hygiene, dressing, ...) and improving child ability to relate with others. The application of this technique at an early age allows to act effectively on the child's behavioral processes, leading to significant results [83]. The ABA therapy has generally three main phases: the first one is merely the observation of child behaviour and reaction to external stimuli. Subsequently, the ABA therapist analyses the behavioural reactions of the child. Finally, the ABA operator draws up a program of specific and personalized exercises to modify the dysfunctional behavior of the child [84].

The child is constantly monitored to check the actual progress and take note, in the form of qualitative scales, of any encountered difficulty, which may require a variation in the ABA program. However, despite its relevance, this monitoring procedure still heavily relies on either direct observation or revision of video recordings by the operators, both coupled with paper-and-pencil-rating scales [85]. This procedure, which includes both the child observation and behaviour evaluation, beside being time-consuming, is qualitative and may be prone to inaccuracies due to operator fatigue.

To attenuate the issue of perspective evaluation, some promising computer-assisted approaches have been proposed in literature. The majority of them is focused on diagnosing ASD. In [86] the authors propose a DL based algorithm that analyses eye movement patterns from video data to discriminate between children with diagnosed

[1] `https://www.aap.org/en-us/Pages/Default.aspx`

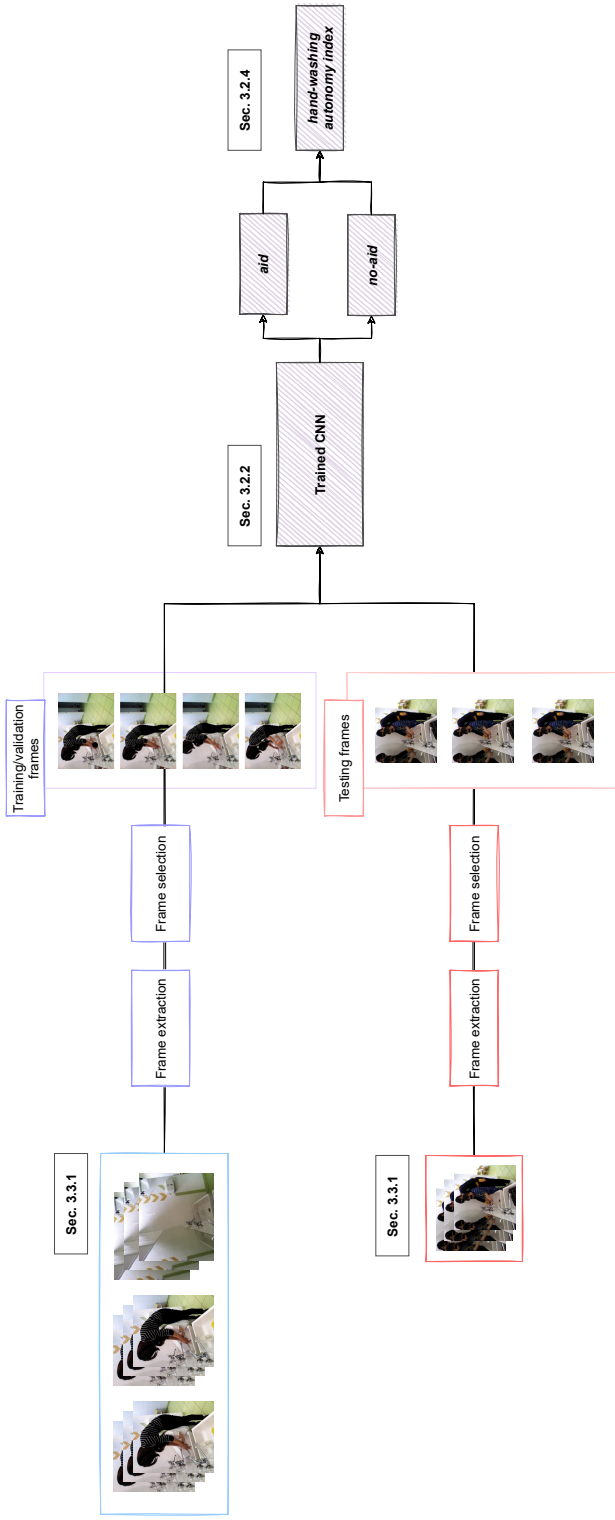


Figure 3.2: Workflow of the proposed deep learning (DL)-based application to classify RGB frames in which the child washes his/her hands with the help of the operator (*aid*) or autonomously (*no-aid*).

ASD and typically developing children. The work in [87] implements a DL framework that analyses video of commonly performed gestures (e.g., grasping a bottle).

Few literature exists on computer-aided systems to support ABA operators in monitoring children with ASD during the therapy session. This may be attributed to the lack of publicly available observational databases. Unlike DL, machine learning-based approaches, require a handcrafted-feature extraction step. This procedure, which could be performed either manually or via specific feature extraction algorithms, may be computational expensive, limiting the translation of such application in the actual monitoring practice. More in general, the use of wearable sensors may alter the behavior of the monitored child, especially for the youngest ones.

To offer all the possible support to the operators in monitoring children with autism during the ABA therapy, this research proposes a DL-based application to analyze images collected from an RGB-D camera. The work elects as case of study the monitoring of hand-washing autonomy. Thus an RGB camera was placed over the bathroom sink (Fig 3.1) of the "Orizzonte centre" (Macerata, Italy) which is specialised in ABA therapy for autism[2]. The camera records the ABA operator intent on teaching the child what to do to wash the hands, autonomously.

The proposed DL algorithm aims to detect, from an RGB frame, whether the child is washing hands autonomously or with the support from the ABA operator. Then, based on the prediction of the algorithm, an intuitive washing-hand autonomy index is calculated.

3.2 Methods

The workflow of the proposed DL approach is showed in Fig. 3.2

3.2.1 Data acquisition protocol: the hand-washing case of study

Personal autonomy skills are certainly one of the elements that mostly affect the quality of life of the child with ASD: being independent from assistance for personal needs might change the future of these children and the way they relate to the environment [88]. This research considers, among the basic autonomies, that of hand-washing which is fundamental for the safety of the person, for ameliorating social integration and to strengthen the child's self-esteem.

As showed in Fig. 3.1, to accomplish the goal, an RGB camera was placed on the corner of the bathroom of the ABA centre to film over the sink. The acquisition set-up, which consisted of an RGB camera (Astra Mini S-Orbbec®) and a minipc Intel® NUC core *i5*, was installed to be imperceptible and to not distract the child during the therapy.

[2] https://www.ilfarosociale.it/

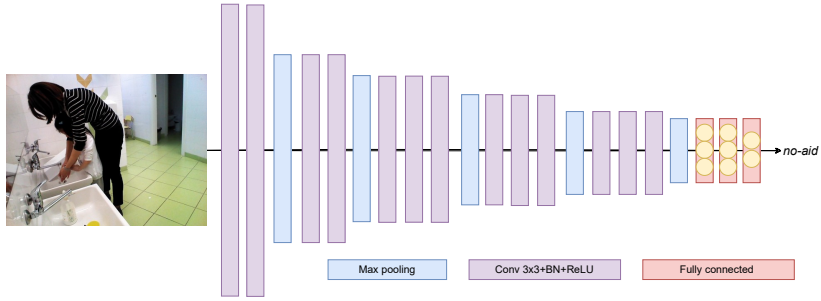


Figure 3.3: VGG16 neural network to classify *aid* and *no-aid* frames. In purple convolutional layers, in blue max pooling layers, in red fully connected layers.

Table 3.1: Dataset description: we used the annotated frames from 36 videos to train and validate the architecture (70% of frames to train and 30% of frames to validate), while the annotated frames of 1 video were used to test the architecture.

Training set		Validation set		Test set	
no-aid	*aid*	*no-aid*	*aid*	*no-aid*	*aid*
3124	3470	1339	1488	201	118
36 videos from 9 children				1 video from 1 child	

A custom-built python script was implemented to automatically acquire concatenated video sequences of 5 minutes each[3]. After gaining the authorization by the children's legal guardians, the acquisition sessions were carried out using a digital programmable timer, for one month, six hours per 5 days (from Monday to Friday).

3.2.2 Network architecture

To classify the selected frames in *aid* and *no-aid* VGG16 network was implemented as a trade-off between low model complexity and good predictive power.

In the original VGG16 implementation [89], the input 224x224 RGB image is processed through 13 convolutional layers which act as features extractors. Each conv block has filters with a quite small receptive field (3×3 pixels) and is activated by a ReLU activation function. Every two or three convolutional blocks (depending on the network depth), max pooling layers are used to progressively reduce the spatial size of the feature map. Max pooling aims to lower the amount of training parameters, to reduce the computational complexity and consequently the risk of overfitting.

The network ends with 3 fully-connected layers with 4096, 4096, and 1000 neurons, respectively, separated by dropouts to reduce the effects of overtraining of the neural network. The last fully connected layer is followed by a softmax layer, used to predict

[3] `https://github.com/roccopietrini/pyOniRecorder`

the probability of the image to belong to each class of the ImageNet dataset[4], the natural-image dataset used to train originally VGG16.

To accomplish the binary classification task, the last fully connected layer (originally with 1000 neurons as the classes of ImageNet) was replaced with a fully connected layer with 2 neurons (Fig. 3.3).

3.2.3 Training strategy

To train the model fine-tuning methodology was adopted. This procedure allows to migrate the knowledge learned by VGG16 during the training on ImageNet to the binary classification, reducing the risk of overfitting as the features extracted from ImageNet database are very generic [90]. To this goal, the weights of the convolutional blocks and the connections between neurons in the first two fully-connected layers were initialized with the weights of ImageNet, while the last fully-connected layer (i.e., the one with two neurons) was initialized with the standard *Glorot* initialization.

3.2.4 Hand-Washing autonomy index

The Hand-Washing autonomy index (HWI) (Eq.3.1) is computed from network predictions as:

$$HWI = \frac{|N|}{|A| + |N|} \tag{3.1}$$

where $A := \{a \mid \text{child is aided in frame } a\}$ is the set of *aid* frames and $N := \{n \mid \text{child is not aided in frame } n\}$ is the set of *no-aid* frames.

This index is provided to ABA operators to quantify the child's level of autonomy during the hand-washing task. Evaluating the trend of this index over time would allow the ABA operators to assess the progress of the child in performing the task.

3.3 Experimental Protocol

3.3.1 Dataset

The dataset used in this work consisted of 115 RGB video sequences of 5 minutes each. The camera frame rate was 30 frames per second with image resolution of 640x480 pixels.

Of 115 total video recordings, only those in which the child and operators were within the camera field of view for at least one frame were selected. This results in a collection of 37 RGB videos. Considering that the hand-washing action is characterized by low dynamics [91], for each of the 37 videos, 1 frame every 6, was extracted. Thus, for each video, 1080 frames were obtained.

[4] http://www.image-net.org/

Table 3.1 summarizes the division of the dataset into training, validation and testing sets. Among the frames extracted from 36 videos, 70% were used to train the network and 30% to validate it. The annotated frames of the remaining video were used to test the network.

Starting from the 1080 frames in each video, frames with no-one in the camera field of view, and with subjects performing actions different than hand-washing (e.g., dish-washing) were manually discarded. Supported by the ABA operators, the remaining frames were manually assigned to the *aid* and *no-aid* class using a custom-built annotator[5]. 10 children and 6 ABA operators took part in the study.

3.3.2 Training settings

Prior feeding the network, the frames were resized to 224 x 224 pixels, for size compatibility with the pretrained network. To train the fine-tuned VGG16 the binary cross-entropy loss was used. The loss was minimized in 50 epochs with SGD as optimizer and a learning rate equal to 0.0005 decayed by a factor of 2 every 10 epochs.

The batch size was set to 64 as a trade-off between memory requirement and training convergence. The best weight configuration among epochs for each model was retrieved according to the highest *Acc* on the validation set. All the analyses were performed using Keras framework on a Intel® Xeon® Silver 4214 CPU @ 2.20GHz with 230 GB of RAM and a NVIDIA® RTX 2080 8 GB RAM.

3.3.3 Ablation study and comparison with other architectures

As an ablation study the performance of the fine-tuned VGG16 was compared against the VGG16 trained from scratch. In the VGG16 trained from scratch the weights of the convolutional blocks were initialized with *He* initialization while the fully-connected layers were initialized with the standard *Glorot* initialization. The performance of ResNet50, both fine-tuned with ImageNet pretrained weights and trained from scratch, was tested too. Considering the structure and the depth of ResNet50 the binary cross-entropy loss was minimized with Adam optimizer setting the learning rate to 0.0001.

For all the architectures, the batch size and the number of epochs, were set to 64 and 50, respectively.

The final model was chosen, among the 4 architectures, as the one with the highest *Acc* in the test set.

3.3.4 Performance assessment

To assess the performance of the CNNs, the classification *Acc*, Precision ($Prec_i$) (Eq. 3.2), Rec_i and f1-score ($f1_i$) (Eq. 3.3) for the i-th class (with $i \in C$: [*aid*, *no-aid*])

[5] https://github.com/roccopietrini/pyMultipleImgAnnot

Figure 3.4: Example of challenging frames: on the left side the child is aided by the operator (blue square), on the right side the child washes his hands autonomously (yellow square).

were computed.

$$Prec_i = \frac{TP_i}{TP_i + FP_i} \tag{3.2}$$

$$f1_i = \frac{2 \times Prec_i \times Rec_i}{Prec_i + Rec_i} \tag{3.3}$$

3.3.5 Performance assessment on challenging frames

To further validate the final classification model, further tests were conducted to assess its performance on challenging frames selected from the original test set. Challenging frames are those in which the child and the operator were close to each other. Samples of challenging frames are shown in Fig. 3.4.

To quantify the proximity between the operator and the child (i.e., how close they are to each other), it was necessary, first, to identify them within the frame. This was done via the FASTER-RCNN detection network, which was pre-trained on the large-scale COCO [6] dataset for natural-image detection tasks. At prediction time, in this work, only the bounding boxes associated with the *person* class were retrieved.

As quantitative index of proximity was then computed the Overlap Ratio (*OR*) among the two bounding boxes (i.e., the child and operator ones). The *OR* was defined as:

$$OR = \frac{area(P \cap K)}{min(area(P), (area(K))} \tag{3.4}$$

where P and K identified the bounding box of the ABA operator and the child, respectively.

Due to the positioning of the acquisition set-up (next to a mirror), before computing the *OR* index, the FASTER-RCNN predictions were post-processed to delete the boxes corresponding to the person detected in the mirror. The Euclidean distance of each up-

[6] https://cocodataset.org/

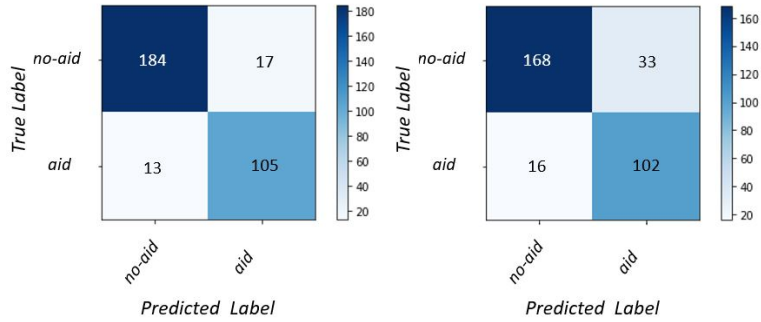


Figure 3.5: Confusion matrix for fine-tuned VGG16 (on the left side) and fine-tuned ResNet50 (on the right side): the two best performing models.

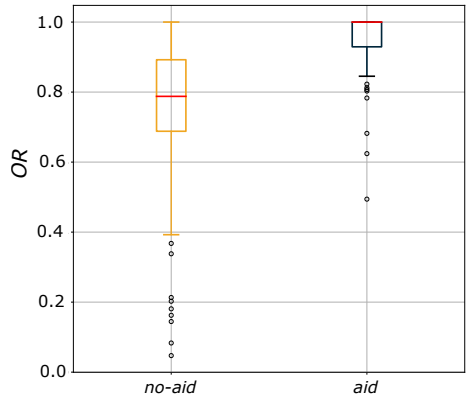


Figure 3.6: Boxplot of the Overlap Ratio (*OR*) for the *no-aid* class in yellow and the *aid* class in blue. Median values of the two distributions are shown in red.

left corner of the predicted bounding box from the origin of the image reference frame was computed. This allowed to select the rightmost bounding boxes excluding those of the mirror.

Then to identify challenging frames the boxplots of the *OR* for the *aid* and *no-aid* class were computed. This procedure allows to exclude all the frames with *OR* lower that the minimum of the boxplot (i.e., the lowest data point excluding any outliers) of the *aid* class.

3.4 Results

Table 3.2 summarizes the results achieved by VGG16 and ResNet50, both fine-tuned and trained from scratch. The two networks trained from scratch achieved the lowest performance when compared with their akin trained with fine-tuning technique,

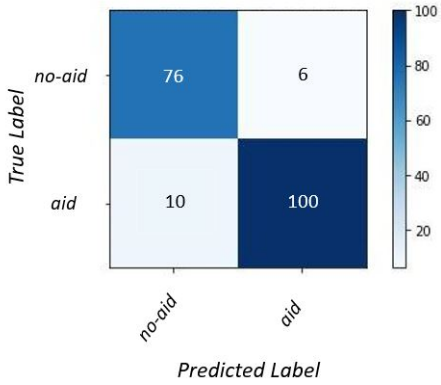


Figure 3.7: Confusion matrix of the fine-VGG16 tested on challenging frames

Table 3.2: Results of the VGG16 and the ResNet50, both with fine-tuning technique and from scratch. Results were evaluated in terms of: class-specific classification precision ($Prec_i$), recall (Rec_i), f1-score ($f1_i$), for $i \in [aid, no\text{-}aid]$ and classification Accuracy (Acc).

	$Prec$		Rec		$f1$		Acc
	no-aid	*aid*	*no-aid*	*aid*	*no-aid*	*aid*	
ResNet50 trained from scratch	0.64	0.63	0.99	0.04	0.77	0.08	0.64
VGG16 trained from scratch	0.76	0.80	0.93	0.50	0.83	0.62	0.77
fine-tuned ResNet50	0.91	0.76	0.84	0.86	0.87	0.81	0.85
fine-tuned VGG16	**0.93**	**0.86**	**0.92**	**0.89**	**0.92**	**0.88**	**0.91**

with extremely unbalanced values of per-class metrics. ResNet50 trained from scratch achieved the worst results with imabalanced values of *Rec* for the *no-aid* and *aid* of 0.99 and 0.04, respectively. VGG16 trained from scratch achieved slightly better results with *Rec* of 0.93 and 0.50 for the *no-aid* and *aid* class, respectively. The results highlighted that both the architectures trained from-scratch are more confident in predicting the *no-aid* class with respect to the *aid* one.

The confusion matrices of the two best performing models (i.e., fine-tuned VGG16 and ResNet50) are shown in Fig. 3.5. Both the models did not outperform in predicting one class with respect to the other. The fine-tuned VGG16 achieved slightly better performance when compared to fine-tuned ResNet50, with higher values of *Prec*, *Rec* and $f1$ for both the *no-aid* class (0.93, 0.92 and 0.92, respectively) and the *aid* class (0.86, 0.89 and 0.88, respectively), with an overall *Acc* equal to 0.91. For the VGG16 the predicted HWI was equal to 0.62 (actual HWI=0.63).

Boxplots of the *OR* for the *aid* and *no-aid* class are depicted in Fig. 3.6. The minimum of the boxplot for the *aid* class (0.82) was used as threshold to select challenging frames. These 192 frames, resulting from the thresholding, were the ones with *OR* greater than the threshold and have been used to further validate the performance of the fine-tuned VGG16. The confusion matrix of the fine-tuned VGG16 tested on challenging frames is shown in Fig. 3.7.

3.5 Discussion

To support the ABA therapists during their actual practice, this research proposed a DL-based application to monitor children with ASD while performing the hand-washing task. By analysing RGB frames, the presented DL model detected whether the child washed hands autonomously (*no-aid* class) or supported by the ABA operator (*aid* class).

VGG16 with fine-tuning technique was implemented as trade-off between model complexity and accuracy in predictions. This model was compared against VGG16 and ResNet50 trained from scratch and fine-tuned ResNet50. Both the performance of the architectures trained from scratch were unsatisfactory as the network poorly predicted the *aid* class. Fine-tuning technique has improved the performance of both the ResNet50 and VGG16 with respect to their corresponding trained from scratch. Hence, fine-tuning allowed to migrate the knowledge of the training on the large-scale ImageNet dataset to the classification task of interest, improving the networks generalization ability. However, fine-tuned VGG16 showed better performance in terms of classification *Acc* and per-class metrics. This may be due to the relatively simple and shallow structure (16 layers) of VGG16 coupled with a small-size dataset.

To further validate fine-tuned VGG16, its performance was tested on the most challenging frames among the testing set. These frames were quantitatively identified as the ones in which the ABA operator and the child were close to each other, even if

the child performed the hand-washing task autonomously. When tested on challenging frames, fine-tuned VGG16 achieved encouraging results in predicting both the *aid* and *no-aid* class (with only 6 out of 76, and 10 out of 100 misclassified frames for the *no-aid* and *aid* class, respectively). This suggested that the network was able to extract information from the frame of interest which did not solely rely on the spatial distance between operator and child, proving to be suitable for the task.

3.6 Conclusion and future perspectives

Autism is a pervasive developmental disorder caused by an altered brain development and can manifest with varying degrees of severity. Children with ASD may mainly develop: (i) problems in social interactions, (ii) repetitive and stereotyped behaviors and (iii) impaired communication skills (verbal and nonverbal). These behavioral alterations appear from the first years of life and persist forever. No definitive cures exist but therapies such as the ABA can help the child in achieving his/her maximum development potential [88].

During the implementation of the ABA program, the operators observe the ASD child and take notes on the progress or difficulties encountered to pursue the program goals. These assessments highly depends on experience of the examiners and are often collected in paper format undermining consultation, longitudinal examination and data-sharing.

Several computer-assisted approaches mostly based on wearable sensors are proposed in literature to overcome these qualitative evaluations and to support ABA operators during the clinical practice. However, body-contact sensors are not always appreciated and accepted by the children and may alter their behavior.

The work here proposed draws from both the experiences of the ABA operators who participated in the development of the research and the limitations of the current literature. It replaces wearable sensors with the same RGB-D camera used to monitor preterm infants in Chapter 2 hidden in a corner of the bathroom and uses its frames to automatically classify whether the child with ASD washes his/her hands autonomously. This monitoring system represents a milestone in the non-invasive estimation of the progress of autistic children. Indeed, besides collecting quantitative data useful in improving the customization of ABA protocols, lays the foundation for largely deploying such systems in centers and at home. This will allow ABA operators to continuously stay up to date on the progress of the children as well as parents to feel almost in touch with the reference center for care of their children.

Autism is a complex syndrome which inevitably impacts those who are affected and their families. There are more children with autism than previously thought because techniques for diagnosing ASD have improved. Advances in diagnosis must therefore be accompanied by improvement in the current systems for monitoring the progress of these children and fully assisting their families during everyday life. To

this goal natural extension of the work presented in this chapter will deal with the comprehensive characterisation of stereotypical motor behaviours. Moreover, the autonomy in pursuing other relevant tasks (e.g., brushing teeth) should be investigated. All these computer-aided solutions will be included in a single framework, to expand ABA-operators' possibilities during their actual practice [92] and to improve these children's and their families' care.

Chapter 4

Estimating human pose from RGB-D images acquired via a smart walker

Silvio Garattini[1] writes in his book "Il futuro della nostra salute": "A macroscopic change that has taken place over the last 40 years is the change in the age of the population: fewer births, around 420000 a year, compared with 600000 deaths, which shifts the age upwards. Life expectancy at birth is increasing: about 81 years for men and as many as 85 for women, with the result that there are more old people than young".

This progressive aging of the population draw attention to the need for rehabilitation services as ways to reduce disability and to live as healthy a life as possible. A rehabilitation that should no more limited to intervention in the acute phases of pathologies as to prevent and limit damage and to preserve independence, but extend to the subsequent phases with tailored follow-ups for consolidating the results achieved and improving independent living.

This scenario calls for finding new treatment modalities and prevention strategies that take into account the specific needs of this fragile population in terms of safety, personalization of treatment and support for well-being. The research described in the following chapter was conducted at the University of Minho (Portugal) and takes up these open challenges in the rehabilitation field. It proposes, for among the first time in the literature, a smart walker-integrated monitoring system. The system, with a view to provide quantitative insights to clinicians in the field, assess the 3D person pose from images collected with two RGB-D cameras mounted over the smart walker. The proposed work, which was born and developed during the first wave of the pandemic, resumes two concepts discussed in Chapter 2 and re-applies them in a different scenario, firstly, that of "pose estimation" here applied to the entire human body and, secondly, the need to design solutions deployable in systems with limited computational resources as a way to fully promote such systems worldwide without

[1] Silvio Garattini is an Italian scientist and pharmacologist, president and founder of the Institute for Pharmacological Research "Mario Negri".

any costs barriers. The present of the proposed system concerns the possibility of establishing personalised rehabilitation plans in hospital scenarios. The future could provide a light-weight system that the patient takes home to continue exercising in a controlled manner but in a familiar environment.

4.1 Background and motivation

Gait and Posture disabilities are common [93, 94], and increasing due to the ageing population and to the global incidence of cardiovascular and/or neurological disorders, such as cerebellar ataxia, cerebral palsy and Parkinson's disease, among others [95, 96]. Along with cognitive impairments, individuals with disability may present a lack of stability, affected motor coordination, poor balance, and muscle weakness, leading to an increased risk of falls and fall-related morbidity [94, 97].

Rehabilitation is traditionally conducted by physicians and therapists over long periods of time, demanding high physical effort from rehabilitation professionals with challenges due to variability in clinical evaluation, while being time-consuming and prone to errors due to clinicians' fatigue [94]. Robotics-based rehabilitation is an evolving area that aims to improve the quality of life of motor-impaired people by providing residual motor skills recovery based on repetitive and intensity-adapted training. Along with assistive devices, smart walkers became a popular choice in the context of gait rehabilitation [98, 94].

An automatic and complete spatio-temporal representation of the patients' configuration in space, based on data gathered from built-in sensors, would be highly desirable to provide personalized care [99, 100]. It would allow the extraction of quantitative parameters to monitor and help rehabilitation professionals evaluate patient improvements, simultaneously serving as a basis for downstream human-robot interactions and user-centered control strategies, adjusting to the patient needs in real-time.

Available smart walker solutions for patient monitoring have focused on extracting narrow aspects, such as specific gait parameters, using specialized hardware and traditional software, with no full-body detection, and presenting fundamental limitations when dealing with non-ideal conditions [101]. Moreover, body detection errors are reported qualitatively, with no general validation scheme, limiting comparison across works [101, 102, 97].

Authors in [103] used ultrasound sensors placed facing each of the patients' legs, to obtain a signal used to measure gait cadence. [104] also predicted this metric, but from force sensors on the handlebars by measuring the force on each handlebar caused by the body sideways displacement. While in [105], the authors have resorted to using laser rangefinders aimed at both legs, to obtain the shank locations on a 2D plane parallel to the ground. Pointcloud data, obtained from a depth sensor pointed at the subjects' legs, was used by [102] along with traditional computer vision techniques (e.g., clustering, Hough transform) to detect the feet locations and knee, hip and ankle

kinematics. In the current version of the ASBGo walker, authors in [97] have also resorted to traditional computer vision techniques based on depth and RGB frames obtained from two independent cameras and systems. One used to monitor the subjects feet and legs and another pointed at the chest for the posture, yielding multiple full body metrics.

Although useful, all the above methods suffer from fundamental flaws when dealing with: non-ideal conditions (e.g., feet occlusions), variable positional offsets depending on body-segments thickness or use of model assumptions, relying on multiple subsystems for the different body parts with no unified full-body approach, and thus not exploiting existing movement dependencies. DL algorithms showed great potential for human pose estimation [100, 106], being capable of providing custom-fit solutions for the problem requirements, with low detection error, robustness to environmental conditions while using cheap consumer hardware [107, 106]. Nevertheless, these require large amounts of labeled data to achieve good performance, which, in the case of 3D keypoint positions, is often not trivial and time-consuming to obtain [100], and most methods were not developed for real-time applications [108, 106, 107].

An automated solution to full-body analysis in rehabilitation scenarios has been proposed in [99]. OpenPose [108] is used to infer the patients' keypoints on videos over a rehabilitation session. Features are then extracted and fed to a regression model to produce gait metrics, with good correlation to physicians' reports. However, the solution does not run in real-time which is crucial for enabling these technologies to be translated in the actual clinical practice.

This research addresses the challenges posed in literature and proposes an innovative solution, based on current advances in human pose estimation using DL approaches. The work implements a non-invasive framework for an accurate, real-time, and lightweight full-body human pose estimation deployed on the ASBGo smart walker [97], using visual information coming from two RGB-D cameras mounted on the equipment.

4.2 Methods

4.2.1 Acquisition set-up

4.2.1.1 ASBGo smart walker

The ASBGo smart walker [97], used in rehabilitation, will run the proposed human pose estimation framework. It is equipped with an Intel NUC-6i7KYK (Intel Corporation, The United States) mini-pc (Intel i7 4-core 2.60GHz CPU, 8GB RAM), responsible for all the high-level algorithms and GUI, while communicating with multiple sensors used for status, patient and environment monitoring, using a ROS 1 [109] messaging interface. During rehabilitation sessions it is tasked with running multiple processes concurrently, and has to keep responsive at all times, constraining the

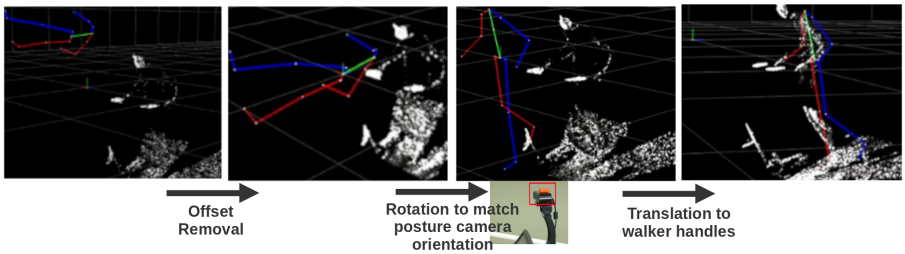


Figure 4.1: Summary of the method used to relate the Xsens keypoint data to the posture camera referential. The skeleton is shown in 3D along with the point-cloud from both cameras.

use of processing-heavy applications. This is challenging considering the lack of any hardware accelerator (e.g., GPU).

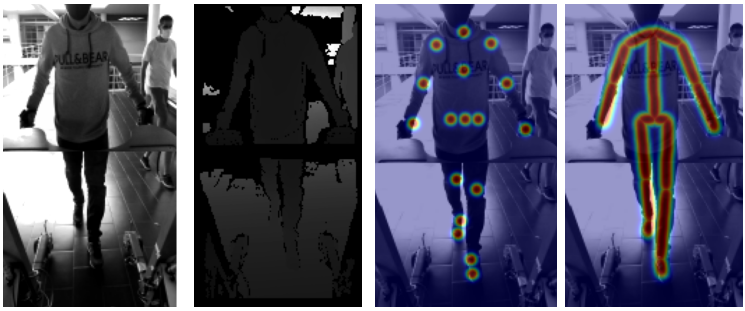
Two Orbbec Astra RGB-D cameras are used to monitor the users and are mounted on the front of the walker, in a configuration with complementing non-overlapping views. The upper camera (posture) only visualizes the upper part of the body, while the bottom camera (gait) only visualizes the legs and feet. Each obtains RGB and also depth images with a resolution of 640x480 pixels at 30 frames per second. The depth sensor has a range from 0 to 10 m (errors increase with the distance from the sensor).

4.2.1.2 Xsens MTw Awinda

The Xsens MTw Awinda inertial MoCap system was used to acquire ground truth data from the subjects using the walker. It is composed of 17 wearable IMU sensors, which communicate over wireless with a base module connected to a computer. The proprietary Xsens MVN software uses the IMU data to drive a biomechanical model of the subject, from which accurate positional and kinematic data are extracted.

4.2.2 Walker dataset

A custom dataset was acquired to train and validate the human pose estimation algorithms: it relates RGB-D images coming from the walker cameras, with ground-truth keypoint data (referred to as skeleton) coming from the Xsens system. A hardware trigger was used to start the acquisition on both systems and the data were later synchronized offline using timestamps saved during recording. The skeleton data were transformed to the posture camera referential. First, the skeleton was centered on the origin of the referential and rotated based on the orientation of an additional Xsens IMU placed on top of the posture camera. Finally, a translation was applied, which places the skeleton wrists on the corresponding walker handles relative to the camera. This translation was obtained through extrinsic calibration using visual markers. The process is summarized in Figure 4.1. With the skeleton aligned in 3D space and by



(a) (b) (c) (d)

Figure 4.2: Processed input image (a) and depth (b) frames which will be fed to the model. stacked Gaussian probability keypoint (c) and connection (d) heatmaps.

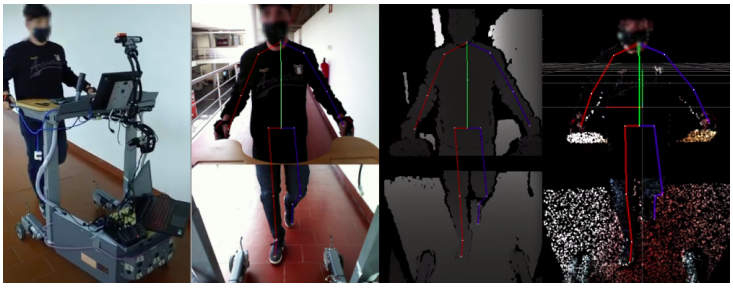


Figure 4.3: From left to right: Outside view of the acquisition setup; collected data from concatenated RGB frames overlaid with projected 2D skeleton; Concatenated depth frames overlaid with projected 2D skeleton; merged point-cloud overlaid with 3D skeleton (data from the gait camera is transformed, through an extrinsic transformation, to the posture camera referential).

knowing the transformation between the cameras, along with the cameras' intrinsics, it is possible to obtain the 2D keypoint locations in each of the camera frames by projecting the skeleton to 2D using the Pinhole Camera model. This process was used to obtain keypoint ground truth labels for the RGB and depth frames for both cameras. The inverse process can also be used to project 2D information from each image to the 3D space.

4.2.3 Dataset preparation

Inspired by [100] all data were down-sampled from the original 30Hz to 10Hz when training the models, to reduce the number of similar samples, which add little additional information to the training set.

4.2.3.1 Input frames

The data used to feed human pose estimation algorithms were pre-processed by: (i) RGB Frames from both cameras were converted to gray scale and normalized to [0,1] range. The depth frames, were divided by the maximum range (10 m), obtaining data between [0,1]. Normalization was preferred over standardization as it preserves the meaning of the depth values. (ii) Posture and gait camera frames were concatenated to create a single frame with information from both parts of the body. This method was chosen over processing frames from each camera individually or doing feature fusion inside the model, as this way positional relationships can be exploited, while decreasing computational costs of processing the two frames independently. This was only possible due to the complementing views obtained from the camera placement on the walker. Some samples of the pre-processed data can be seen in Figs. 4.2a, 4.2b. (iii) The frames resolution was reduced, decreasing computation and memory requirements and thus inference time. Since the subject will always be close to the camera while using the walker, the loss of fine details should not cause a decrease in performance, while on the other hand, increasing the percentage of the frame present in the effective receptive field (ERF) of the model, without requiring a very deep architecture. Finally, 2 input features, one for the depth and another for the grayscale image, with shapes (1, H, W) were obtained.

4.2.3.2 Keypoint selection

A subset of 17 keypoints was selected from the original Xsens skeleton obtained in the data acquisition step. A set of 16 connections were also defined between these keypoints, following natural limb segments of the skeleton. The keypoints and connections can be seen in Figure 4.3.

4.2.3.3 Heatmaps

Gaussian probability heatmaps were created from the 2D keypoint locations. These will be used as intermediate regularization for the model, following the human pose estimation literature [100, 110, 111]. Keypoint heatmaps were created for each of the keypoints, by a Gaussian probability function centered at each 2D location and with a variance (σ) of 3 [111]. Similarly, connection heatmaps were created between connected keypoints by using 1D Gaussian distributed values along a connection line between two keypoints. The stacked heatmap output features for each type can be visualized in Figure 4.2b). This creates data samples with shapes (17, H, W) and (16, H, W) respectively for the keypoint heatmaps and connection heatmaps.

4.2.4 Model framework

The DL model framework for human pose estimation is shown in Figure 4.4 and follows a two-stage approach. The 2D-stage is responsible for detecting keypoints and keypoint-connections in 2D image space. The 3D-stage regresses the keypoint locations to 3D space. This two-stage architecture was chosen above methods which directly compute the positions in 3D space [112, 106], as it is lighter to compute and has shown competitive results in the literature [113], while being easier to optimize since there is no internal conversion between image and cartesian spaces [112]. Moreover, it allows dealing with the multi-camera fusion problem in separate stages without having to learn a global internal representation.

4.2.4.1 2D-stage

The 2D-stage of the model is responsible for detecting the keypoint-pixel locations from the input image and depth frames. This is accomplished, inspired by literature in closer fields [108, 110, 114, 115, 107, 116, 117], with a fully convolutional network. Special attention was given to the computational cost given the hardware constraints. The architecture is displayed in Figure 4.5.

A backbone based on the lite variants of the EfficientNet [118] architecture was used as a general feature extractor similarly to [107], settling for the lite0 version. The first 4 resolution feature blocks were selected, yielding multi-level features, with dimensions from $(\frac{W}{2}, \frac{H}{2})$ to $(\frac{W}{16}, \frac{H}{16})$.

Two decoder branches, with skip connections from the backbone intermediate features [110], up-sample these representations to produce the output keypoint and connection heatmaps respectively on each branch, with $(\frac{W}{2}, \frac{H}{2})$ resolution. The bottleneck features $(\frac{W}{16}, \frac{H}{16})$ were enhanced using an atrous spatial pyramid pooling module (ASPP) [119], to increase the ERF of the model, similarly to [117].

The bias of the last convolution layers before each output heatmap is initialized with the method proposed by [120]. This decreases the probability of keypoint detection

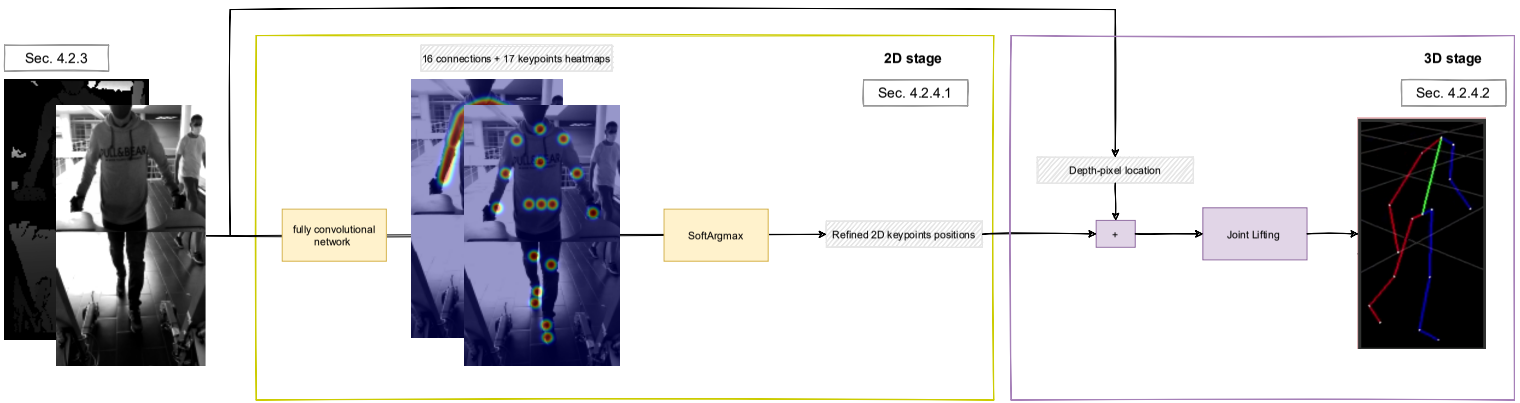


Figure 4.4: Proposed two-stage model framework. The 2D-stage takes the input frames and regresses the keypoints and connections heatmaps using a fully convolutional network. Soft-argmax is used on the keypoint heatmaps to obtain the 2D keypoint locations, which are lifted to 3D space using a fully connected regression network aided by the depth information.

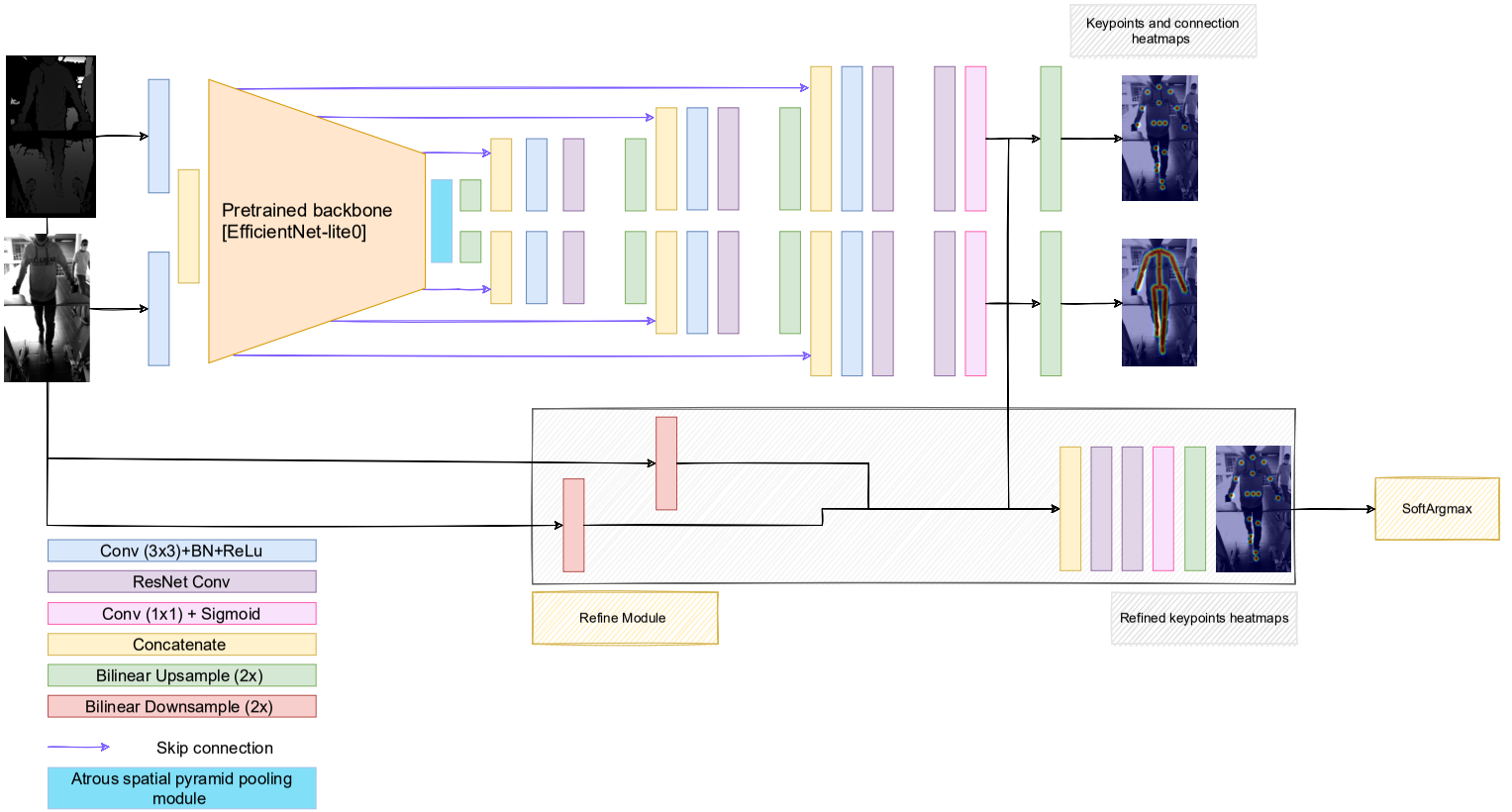


Figure 4.5: 2D-stage model architecture: The processed input image and depth frames are concatenated and passed by a pretrained backbone (EfficientNet-lite0), yielding multi-level features. The higher level features are enhanced by an atrous spatial pyramid pooling module and then up-sampled through two branches, to produce connection and keypoint heatmaps. A refining module is used to improve keypoint heatmaps, from which the joint locations are extracted using the soft-argmax operator.

in each pixel, reducing early training instability due to the inherent class imbalance associated with the sparse heatmap values.

A shallow refining module is used to improve the keypoint heatmaps, by aggregating information from both heatmap branches. The computation is performed in half resolution size to decrease the computational cost, and the output heatmaps are upsampled to the input frame size through bilinear interpolation, following [110, 117].

Finally, each keypoint location in the input frames is extracted from the refined keypoint heatmaps using the soft-argmax operator [116], along with the detection confidence.

4.2.4.2 3D-stage

The 3D-stage deals with the residual linear model proposed by [113], lifting the 2D keypoint locations from the previous stage to 3D. The 2D keypoint locations are previously normalized, projecting them from pixel locations to normalized pixel space, while maintaining the aspect ratio. This increases training stability and generability to different input frame resolutions. Additionally, depth information at each pixel location was given to help resolve pose ambiguities and ease the learned problem of point projection.

A total of 256 neurons per hidden layer were used, instead of the original 1024 neurons [113]. In a preliminary analysis, this was found to reduce overfitting while making the model faster to train. This may be explained given the smaller number of parameters (roughly 0.29 M for 256 neurons compared to 4 M for 1024 neurons compared to 0.29 M for 256 neurons).

4.2.4.3 Losses

The 2D-stage of the model was trained using the integral loss proposed by [116], where the mean absolute error (*MAE*) between the predicted 2D keypoint positions and the corresponding ground truth is minimized. This is combined with a heatmap *MSE* regularization term applied to all heatmaps.

As in [121], the 3D-stage was trained with a Log-Cosh loss between 3D keypoint positions and corresponding ground truth. This loss combines the outlier robustness of the *MAE* with the diminished update size for smaller error of the *MSE* loss.

4.2.5 Deployment

After offline training, the model was exported to *.onnx*. This removes most dependencies while offering optimization tools to improve runtime latency. The model was loaded to the walker existing C++ ROS 1 (Melodic Morenia) [2] environment using the *.onnx* accelerator library built with the default CPU provider.

[2] `http://wiki.ros.org/melodic`

4.3 Experimental protocol

4.3.1 Dataset details

Trial conditions followed standard gait rehabilitation procedures with the walker, and were defined in collaboration with clinicians. Each subject was instructed to perform 27 trials composed of: 3 sequences (walking forward, cornering left, cornering right); at 3 different speeds (0.3, 0.5, 0.7 m/s) [122]; each repeated 3 times (with the same course) in different corridors (to maximize environment variability. The final dataset contains a total of 378 trials, from 14 healthy subjects.

This amounts to 166k frames of synchronized data sampled at 30 Hz (92 minutes of total recording time). Data from each subject were divided into train (9 subjects), validation (2 subjects), and test (3 subjects) splits, resulting in $\sim$110 k, $\sim$19 k, $\sim$36 k samples for each split, respectively. A sample can be seen in Figure 4.3.

4.3.2 Implementation Details

The image and depth input frames were resized from a resolution of 640x960 pixels[3] to 128x224, decreasing computation and memory requirements and thus inference time.

Models were trained in two steps. First, only the 2D-stage was trained, using the image and depth frames as inputs to produce the intermediate 2D features: keypoint locations and the keypoint and connection confidence heatmaps. The 3D-stage of the model was then trained by using as input the 2D features, with that stage weights frozen (to prevent destroying previously learned features), and outputting the 3D keypoint positions. End-to-end learning of the complete model was tried in early experiments, however, it led to worse 2D heatmaps and was thus dropped, also decreasing training complexity.

Reasonable hyper-parameters for training were found empirically and kept constant for all models tried. The Adam optimizer was selected with an initial learning rate of 0.002 which decayed to 0.00001 over 30 epochs using a cosine-annealing schedule. A batch size of 16 and 32 was, respectively, used to train the 2D and 3D-stages of the model. Gradient clipping with a range of [-0.2, 0.2] was applied during training for all models to prevent high gradient updates, especially on the first batches of training which could destroy some of the pre-trained weights. All convolutional and fully connected layers, except for the outputs, were followed by batch normalization and a ReLU to activate.

Random train-time data augmentation [123] was used to increase visual variability of the training set, decreasing over-fitting to the limited number of train samples. However, the depth frames could not be augmented with common operations (as it would produce incorrect depth information), so only pixel dropout was applied. Occlusion

[3]Concatenated frames have a height of $2 \times 480 = 960$ pixels.

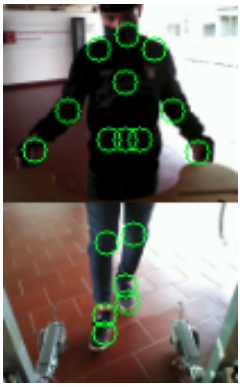


Figure 4.6: Chosen percentage of correct keypoints (*PCK*) threshold radius of 6 pixels. Detection values inside the circles, for each keypoint, are considered correctly detected.

specific augmentation was also applied following the method by [124] of adding random occlusion objects to the image frames, with dropped depth pixels. Additionally, when training only the 2D-stage, affine transformations were applied to decrease over-fitting to frequent keypoint locations on the walker dataset. This produced slight incoherence in the depth information, but led to better overall results.

Additional train-time regularization was applied to most layers of the model in the form of dropout, with a percentage of 50% for all the linear layers and spatial-dropout with a probability of 20% for the convolutional layers. A L2 weight decay parameter with a value of 0.00001 was also added.

4.3.3 Performance metrics

The following metrics were used to assess the models performance:

- Mean per-joint position error (*MPJPE*): average Euclidean distance between a ground truth position and a predicted position for each of the keypoints is calculated after performing root joint alignment (in this case the pelvis). It will be the most focused in this work, since most downstream patient analysis applications use root-relative joint positions.
- Procrustes-aligned *MPJPE* (PA_*MPJPE*): it was first permed a Procrustes analysis ignoring affine errors then MPJPE is computed.
- Absolute *MPJPE* (*A_MPJPE*): similar to *MPJPE* but uses absolute positional values relative to the camera.
- Percentage of correct keypoints (*PCK*): percentage of predicted keypoints with an error below a certain threshold. As in [100] 75mm was selected for all 3D

tests. This metric was also evaluated in 2D where a threshold of 6 pixels was selected and can be seen in Figure 4.6.

- Inference time (latency): it was evaluated on a Nvidia Tesla T4 GPU and on an Intel i7 - 4720HQ CPU. This is critical given the real-time nature of the application, and refers to the time necessary to load the inputs into the device and do a forward pass for a single sample. The results were obtained by running inference on all samples in the test split.

4.3.4 Model variants

Alternative versions of the 3D-stage were also considered, based on ideas presented in the literature, and in a search for the best possible configuration.

4.3.4.1 Baseline

This variant implements the original model proposed by [113]. It is similar to the default 3D-stage method but without including the depth information as input to perform the lifting and considering the original 1024 channels per layer.

4.3.4.2 Semantic graph convolutions

Semantic Graph Convolutions network (SemGCN) [125] was tested, with the addition of depth information for each keypoint as input. This module exploits the hierarchical structure of the skeleton by using state-of-the-art graph convolutions that aggregate information along connected joints. The non-local layers used in the original work were not used here as they doubled the inference time without noticeable improvements.

4.3.4.3 Projection residual

This method follows the approach by [121], extending it to a non-overlapping multi-camera setup. Instead of learning to project the data from 2D space to 3D similarly to the lifting methods, an explicit projection is computed and then refined, as follows.

The detected 2D keypoints in each of the camera frames are first projected to 3D space relative to the posture camera referential, using the depth information and the intrinsic parameters for each of the cameras, based on a Pinhole camera model. Then it is applied an extrinsic transformation to the projected gait camera data, so it is in the posture camera referential, this produces a rough estimate of the detected keypoint positions, but affected by fundamental flaws: (i) homogeneous pixels intensity levels which occurs quite frequently, and make it impossible to project the affected keypoints to 3D space; (ii) the person's body thickness, which will produce a varying offset for each keypoint and will be highly dependent on the subject using the equipment; (iii) projection of incorrect pixels due to bad detections in 2D space, which can

lead to unreasonable keypoint locations, especially when this occurs for points in the background; (iv) cannot deal with keypoint occlusions, since these would be projected incorrectly to 3D space; (v) small random error dependent on the depth sensor noise, which reduces temporal coherence.

The 3D flawed transformation can be improved by applying a series of residual fully connected layers to produce a globally coherent result by using spatial information from the other keypoints.

This method provides a more principled way to fuse the depth information from both cameras. An explicit transformation (which can be obtained through extrinsic calibration with low error) is used instead of relying on a learned internal representation that would need retraining for different camera spatial arrangements.

4.3.4.4 Spatio-temporal feature analysis

A sequential frame approach was also tried. It aggregates temporal information from multiple frames, as the one proposed in Sec. 2.3 and is achieved by stacking successive frames (W_d= 4 frames) which are processed using temporal convolution blocks. These not only aggregate local spatial information, but also temporal information.

All received frames are processed simultaneously, yielding corresponding predictions for each one. In the case of the 2D-stage, the 2D convolutions are replaced with 3D convolutions, while in the case of the 3D-stage, the fully connected layers are replaced with 1D convolutions, where the extra dimension represents time.

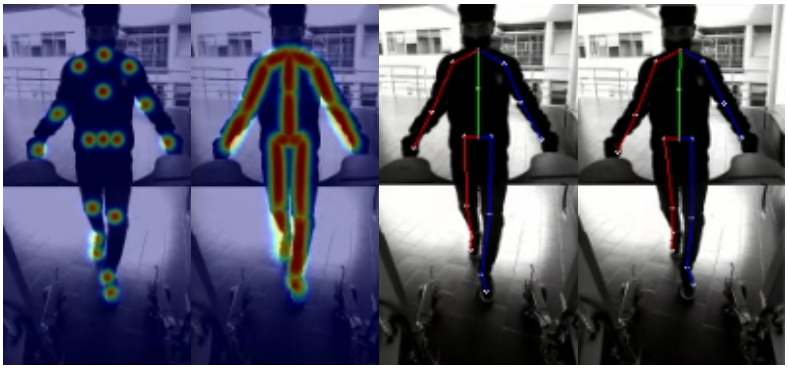
The 2D-stage backbone was also replaced with a similar temporal MobilenetV2 [126] architecture (since no pre-trained temporal EfficientNet version was found). The low-level layers also have temporal pooling operations removed (spatial pooling is still performed), as temporal invariance is not desirable since a sequential prediction for all 4 frames is necessary.

4.4 Results

4.4.1 2D-stage

The 2D-stage (Section 4.2.4.1) was evaluated in the following section. Examples of features obtained for a frame of one of the test subjects can be seen in Figure 4.7. The model not only produces the 2D keypoint locations, but also the refined keypoint heatmaps used to obtain them, and the connection heatmaps, these are compared against the ground truth skeleton from the Xsens.

The detection error for each keypoint was further analysed through a boxplot graph depicted in Figure 4.8. All keypoints display a relatively similar detection error, with a mean of 3.76 pixels, corresponding to 85.27% of detections being below the chosen detection threshold of 6 pixels, with an inference time of 11.97 ms and 37.56 ms in the GPU and CPU respectively.



(a) (b) (c) (d)

Figure 4.7: **(a)** Keypoint heatmaps, **(b)** Connection heatmaps and **(c)** 2D keypoints and connections predicted by the 2D-stage of the model. **(d)** Corresponding 2D keypoint ground-truth labels. All data are overlaid on top of the input image frame.

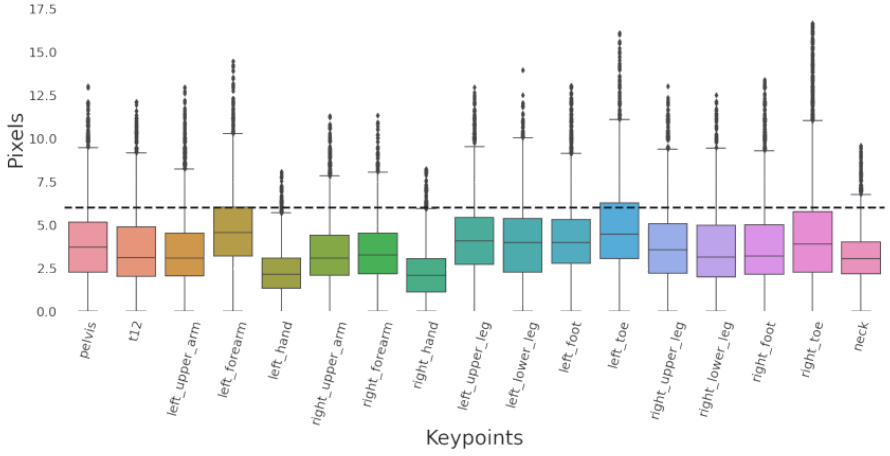


Figure 4.8: Boxplot per-joint error (*MPJPE*) for the 2D detection, across all test frames obtained from the 2D-stage (extreme outliers were removed for better visibility). The dashed line marks the 6 pixel threshold defined.

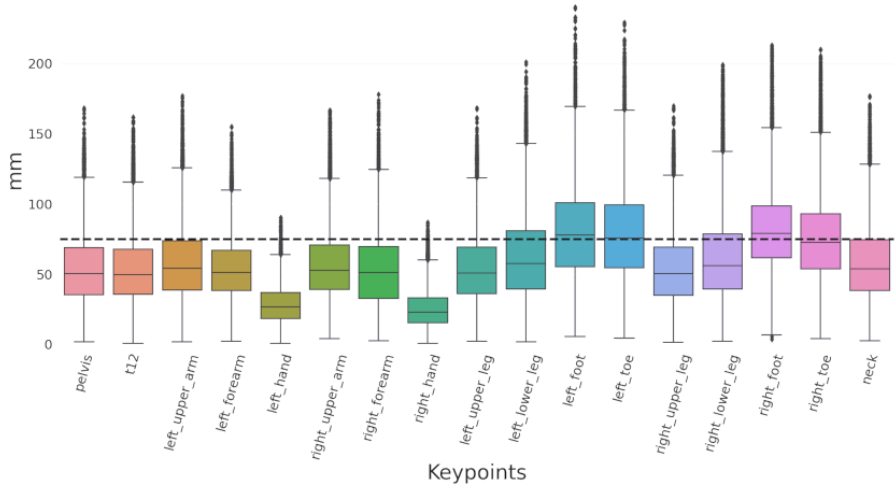


Figure 4.9: Boxplot per-joint error results for the 3D detection, across all test frames, with absolute values relative to the posture camera referential, obtained with the complete model (extreme outliers were removed for better visibility).

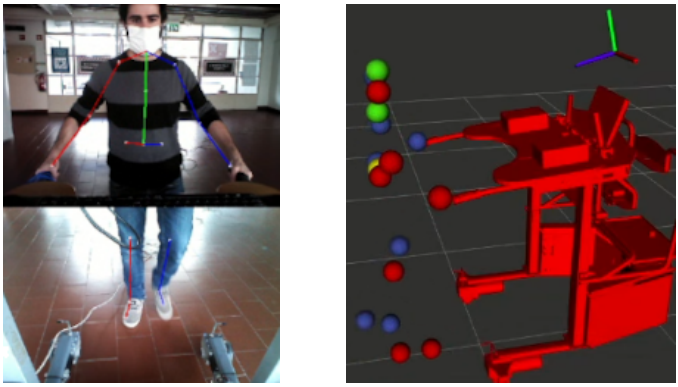
Table 4.1: Results summary of the 3D-stage and comparisons against the different variants for regression from the 2D-stage keypoints. The best results are highlighted in bold.

Method	*MPJPE*[mm]	PA_*MPJPE*[mm]	*PCK*@75[%]	GPU latency[ms]	CPU latency[ms]	Parameters
2D-stage						
+ Default 3D-stage	**44.05** ± 0.39	**33.35** ± 0.29	**83.03** ± 0.48	**13.43** ± 0.03	**38.93** ± 0.13	1.34 M
+ Baseline	45.80 ± 0.40	34.78 ± 0.31	81.31 ± 0.5	14.68 ± 0.04	41.28 ± 0.10	5.34 M
+ SemGCN	45.85 ± 0.42	34.80 ± 0.31	81.45 ± 0.51	19.21 ± 0.03	43.86 ± 0.15	**1.31 M**
+ Projection_Residual	48.35 ± 0.44	36.76 ± 0.39	79.88 ± 0.49	14.50 ± 0.02	39.18 ± 0.12	1.34 M

4.4.2 Complete model

Figure 4.9 depicts the 3D error for each keypoint for the complete model (Section 4.2.4) relative to the camera. An average absolute error of 59.5 mm relative to the posture camera was obtained. A root-relative error of 44.1 mm was obtained, with a *PCK* of 83.0%. The feet keypoints displayed the largest mean errors, closer to the imposed detection threshold. After applying a Procrustes transformation, the error is around 33.3 mm, signaling the presence of affine errors, in the form of positional offset, rotation or scale, which when removed yield a *PCK* of 96.3%.

The default model was further compared with other 3D-stage alternatives explored in Sec. 4.3.4 from the predictions of the default 2D-stage. The results are summarized in Table 4.1. A similar performance was encountered for all models with a *MPJPE* $\sim$ 46.0mm and *PCK* of $\sim$ 81.4 mm.



(a) (b)

Figure 4.10: Predictions obtained from the model running on the smart walker, in 2D (a) and 3D (b) spaces. Connections between the hips and legs are not rendered in 2D due to the discontinuity between camera frames. The 3D visualization used the RViz package from the ROS environment to render the 3D keypoint locations relative to the walker. The posture camera frame of reference is also displayed.

4.4.3 Deployment

The *.onnx* Runtime optimizations decreased the runtime latency of the model from the original 38.9 ms to 23.1 ms, with an additional 3.5 ms to pre-process the input frames (Section 4.2.3.1), while keeping a similar detection error.

During a normal rehabilitation session, due to other concurrent systems running, the latency of the human pose estimation framework integrated on the ROS environment was higher, with also higher variability, ranging from 25 ms to 70 ms (averaging 40 ms), with a mean of 40 ms. Moreover, results are published with some response delay (around 0.3 s) given the asynchronous nature of the underlying ROS system in the equipment.

During normal walking the model performs well, being capable of detecting the 2D and 3D keypoint locations (Figure 4.10) as expected. However, it performs sub-optimally when confronted with body configurations that lie outside the normal training distribution (e.g., when walking on the sides of the walker, bringing the feet high above the ground, complete feet occlusion).

4.4.4 Benchmark and ablation studies

Multiple ablation studies were conducted by removing certain components of the complete pipeline, to identify the contribution of each to the overall results. The findings are described in the next section, along with a comparison to alternative model variants (Sec. 4.3.4).

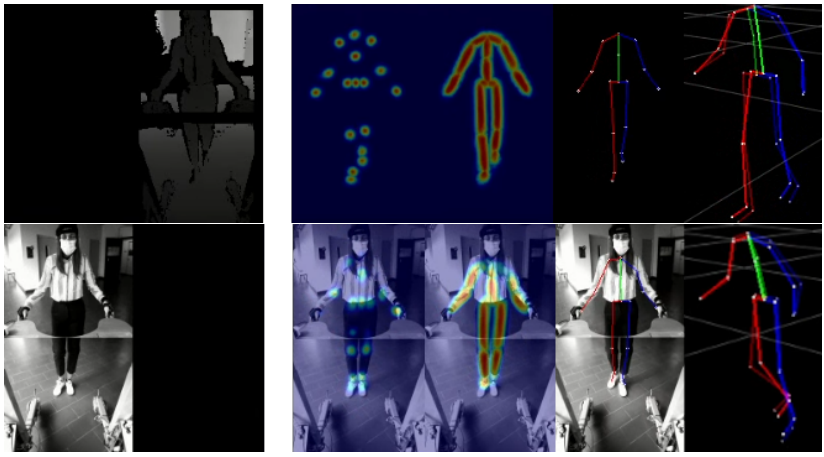


Figure 4.11: Results obtained when removing information from image (upper images) and depth (lower images) inputs. The image+depth inputs for each experiments are grouped in the left side, while the corresponding model predictions (keypoint heatmaps, connection heatmaps, 2D keypoints, 3D keypoints) are grouped in the right. 2D outputs were overlaid on the image input. 3D ground truth keypoints are shown overlaid with higher transparency.

4.4.4.1 Input modalities

The proposed model is not too dependent on any of the two input modalities (image, depth). It is robust to corruption of one of the input frames, giving reasonable predictions by focusing on information from the other, even though not being explicitly trained to do so. Examples of this can be seen in Fig. 4.11.

The model appears to be more affected by corruption of the depth input, giving low confidence detections on the heatmaps with also higher degree of keypoint error in the 2D space. Nevertheless, even without the depth information, it is still capable of regressing the positions of the 3D keypoints.

On the other hand, it seems to be less affected by corruption in the input image, as the heatmap predictions still display a well-defined shape, with high detection confidence.

4.4.4.2 Backbone

The chosen EfficientNet 2D backbone performance was benchmarked against commonly used ResNet architecture in Table 4.2. All models achieved a similar detection error around 3.75 pixels, with slightly better results for the ResNet50 backbone ($MPJPE\_2D$= 3.66). However, it displayed a significantly larger latency (238.9 ms), being 6.36 time slower when processing in CPU compared to the default EfficientNet-lite0 option (37.56 ms), which displayed the best computation time in the CPU, being

Table 4.2: The default EfficientNet-lite0 backbone 2D-stage performance in comparison with the common ResNet models. OpenPose is also compared in terms of latency for reference (values were taken from the paper), as it is a common baseline on real-time huma pose estimation. The best results in each metric are highlighted in bold.

Architecture	*MPJPE*_2D [pixels]	*PCK*_2D@6[%]	GPU latency[ms]	CPU latency[ms]	Parameters
Default	3.73 ± 0.04	85.27 ± 0.59	11.97 ± 0.03	**37.56** ± 0.13	**1.05** M
ResNet50	**3.66** ± 0.04	**86.93** ± 0.56	15.27 ± 0.04	238.89 ± 0.66	26.8 M
ResNet18	3.88 ± 0.05	84.65 ± 0.61	**11.88** ± 0.03	94.70 ± 0.35	12.1 M
OpenPose[108]	-	-	∼36.00	∼10396.00	-

Table 4.3: Importance of the refine module on the 2D-stage accuracy and latency. Its effect is investigated by removing the refine module and obtaining the 2D keypoint locations from the keypoint heatmaps branch, and by further removing the parallel connection heatmap branch. The best results in each metric are highlighted in bold.

Architecture	*MPJPE*_2D[pixels]	*PCK*_2D@6[%]	GPU latency[ms]	CPU latency[ms]	Parameters
Default 2D-stage	**3.73** ± 0.04	**85.27** ± 0.59	11.97 ± 0.03	37.56 ± 0.13	1.05 M
- Refine Module	4.37 ± 0.10	81.92 ± 0.71	11.02 ± 0.02	31.90 ± 0.10	1.0 1 M
- Connections Branch	4.37 ± 0.10	81.92 ± 0.71	**9.07** ± 0.04	**25.89** ± 0.07	**0.9** 3 M

faster than the light ResNet18 (94.7 ms) by 2.5 times while slightly more accurate.

The commonly used OpenPose [108], was also considered for the 2D-stage, as it boasts good performance with real-time inference. However, this option was quickly dropped after checking the latency on the CPU, where the authors point to a latency around 10 s for a single frame, making it unacceptable for real-time human pose estimation on the walker hardware.

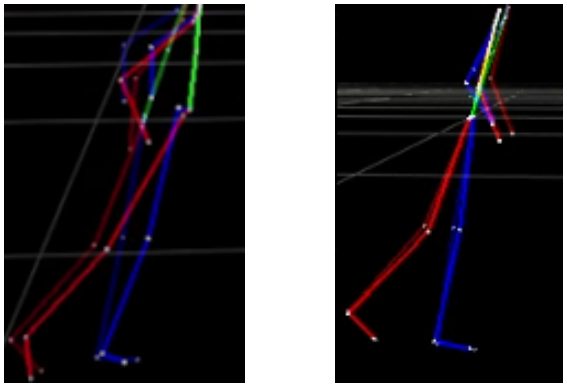
4.4.4.3 Refine module

Only the keypoint heatmaps are necessary to extract the keypoint locations, which can be extracted directly from the heatmap branch in the case where the refine module is not used. Thus, it is possible to completely ignore the computation of the connection heatmaps branch to increase run-time performance. Table 4.3 compares the detection results of the keypoint 2D locations with and without using the refine module with the default 2D-stage. Additionally, are also presented the results after removing the parallel connection heatmaps branch entirely.

Removing the refine module yields a 17% increase in the *MPJPE* error which is traded for an also 17% decrease in latency, which is improved further by 45% after removing the connection branch altogether.

4.4.4.4 Projection residual

Figure 4.12 shows the results obtained from directly projecting the 2D keypoint locations with depth information to 3D space using the camera Pinhole model and sub-



(a) (b)

Figure 4.12: (a) 3D skeleton obtained through keypoint depth projection using the camera Pinhole model (the leg/feet keypoints were transformed to the posture camera referential using the known extrinsic transformation). (b) The skeleton obtained after the residual correction step of the Projection_Residual method.

Table 4.4: Default 3D-stage lifting approach comparison with the 2D keypoint locations projection using the pixel depth and the camera Pinhole model (Projection_Raw), and by further processing using the residual correction neural network in the Projection_Residual method. The ground truth 2D keypoint locations were used as input to the regression in all methods. The best results in each metric are highlighted in bold.

Architecture	*MPJPE*[mm]	PA_*MPJPE*[mm]	*PCK*@75[%]	CPU latency[ms]	Parameters
Default 3D-stage	**36.65** ± 0.28	**28.11** ± 0.24	**90.64** ± 0.34	**1.03** ± 0.0	0.29 M
Projection_Raw	226.0 ± 4.82	179.97 ± 3.11	21.84 ± 0.33	1.13 ± 0.0	**0**
Projection_Residual	43.42 ± 0.41	33.56 ± 0.37	84.72 ± 0.45	2.27 ± 0.0	0.29 M

sequently applying the residual correction. Table 4.4 compares these results against those obtained with the default 3D-stage. The ground truth 2D locations were used to decrease the noise in the results from the 2D-stage error.

The raw projection method displays a larger error, above 200 mm, with only 21.8% of predictions having an error below the desired threshold of 75 mm. The residual correction is capable of improving the detection substantially, to an error around 43.4 mm with 84.7% of keypoints within the detection threshold, although still higher than the lifting approaches.

4.4.4.5 Only 3D-stage

The results from the 3D-stage were evaluated independently from the errors of the 2D-stage, by using the ground truth 2D keypoint locations as input for the regression. These results provide the maximum performance achievable given an ideal 2D-stage

Table 4.5: Result summary of the default 3D-stage and lifting variants when receiving as input the ground truth 2D keypoint locations. The default 3D-stage error when regressing from the 2D-stage predictions is shown as reference. The best results in each metric are highlighted in bold.

Architecture	*MPJPE*[mm]	PA_*MPJPE*[mm]	*PCK*@75[%]	CPU latency[ms]	Parameters
2D-stage					
+ Default 3D-stage	44.05 ± 0.39	33.35 ± 0.29	83.03 ± 0.48	38.93 ± 0.13	1.34 M
2D-ground truth					
+ Default 3D-stage	36.65 ± 0.28	28.11 ± 0.24	90.64 ± 0.34	**1.03** ± 0.0	0.29 M
+ Baseline	36.66 ± 0.28	26.91 ± 0.23	90.75 ± 0.34	3.98 ± 0.0	4.29 M
+ SemGCN	**35.22** ± 0.28	**26.25** ± 0.25	**92.59** ± 0.34	4.38 ± 0.0	**0.27M**
+ Projection_Residual	43.42 ± 0.41	33.56 ± 0.37	84.72 ± 0.45	2.27 ± 0.0	0.29 M

Table 4.6: 2D-stage results obtained using the temporal model with 4 sequential frames, compared to the default single-frame version.

Architecture	*MPJPE*_2D[pixels]	*PCK*_2D@6[%]	GPU latency[ms]	CPU latency[ms]	Parameters
Single(1 frame)	**3.73**± 0.04	**85.27** ± 0.59	**11.97** ± 0.03	**37.56** ± 0.13	**1.05M**
Sequential(4 frames)	4.13 ± 0.06	83.41 ± 0.63	51.13 ± 0.05	554.08 ± 2.07	1.40M

and allow a less noisy comparison. Table 4.5 depicts the results and the comparison to the complete model.

In general, all lifting methods displayed a similar level of performance (around 36 mm *MPJPE*) with a low latency time (<4.38 ms). The SemGCN variant obtained overall the best results, while the projection residual approach obtained once again the largest error.

An almost 8mm higher *MPJPE* (36.65 mm) was obtained when lifting from the 2D-stage keypoints with the default model, compared to using the 2D ground truth keypoints (44.05 mm).

4.4.4.6 Temporal Model

The results obtained using the spatio-temporal stages (Section 4.3.4.4) were compared next, starting with the 2D-stage in Table 4.6 and the 3D-stage (from 2D keypoint predictions and ground truth in Table 4.7.

The 2D spatio-temporal model obtained slightly worse detection results, with a 2D *MPJPE* of 4.13 pixels, while being noticeably slower to compute, especially on the CPU, taking 554 ms to process 4 frames, compared to the single-frame counterpart.

The 3D-stage yielded similar performance to the default single-frame version, being marginally worse when using the predictions from the 2D-stage, given the slightly worse results obtained by the sequential 2D-stage. The latency of the temporal 3D-stage is slightly higher than the single-frame model but it is overall faster since 4 frames are processed simultaneously.

Table 4.7: 3D results obtained using the temporal model with 4 sequential frames, compared to the proposed single-frame version, from the 2D keypoint location predictions of the 2D-stage and ground truth. The best results in each metric are highlighted in bold.

Architecture	*MPJPE*[mm]	PA_*MPJPE*[mm]	*PCK*@75[%]	GPU latency[ms]	CPU latency[ms]	Parameters
2D-stage						
+ Single(1 frame)	**44.05** ± 0.39	**33.35** ± 0.29	**83.03** ± 0.48	**13.43** ± 0.03	**38.93** ± 0.13	**1.34M**
+ Sequential(4 frames)	46.72 ± 0.56	34.93 ± 0.41	82.61 ± 0.51	52.50 ± 0.06	554.86 ± 2.45	1.98M
2D-ground truth						
+ Single(1 frame)	36.65 ± 0.28	**28.11** ± 0.24	90.64 ± 0.34	**1.31** ± 0.00	**1.03** ± 0.0	**0.29M**
+ Sequential(4 frames)	**36.49** ± 0.29	28.79 ± 0.23	**91.01** ± 0.36	1.77 ± 0.00	1.60 ± 0.0	0.58M

4.5 Discussion

This research proposes a real-time full-body pose estimation solution for the ASBGo smart, able to extract a compact body configuration representation, from the RGB-D cameras stream, which can be used as prior for extracting multiple gait. Multiple benchmarks and ablations studies were performed to evaluate and better understand the model performance, as well as exploring competing approaches.

The ground-truth labels obtained through the acquisition method are not perfect since the Xsens data contain no visual correlation with the camera streams during acquisition, resulting in positional offsets, which is worsened by compounding errors when relating both the referentials (Xsens calibration errors, extrinsic camera calibration, users wrist position on the handles). These errors can result in bad samples which do not align visually with the limbs on some frames.

Depth data could be used to pre-process the frames to segment the user by applying geometric and threshold operations, before feeding them to the models as in [102]. However, this approach was not followed since the neural networks are capable of parsing this information from the raw frames, the additional processing overhead, and the introduction of failure cases which would decrease the models' robustness.

The location objective is rather ambiguous below a certain threshold, since many locations in the neighborhood could be considered correct for each keypoint. Moreover, due to the presence of some noise in the ground truth labels, it would be impossible to obtain no error. The wrist keypoints displayed the lowest errors, given the low variability as the subjects are required to be grabbing the walker handles at all times, while, the feet keypoints displayed the largest errors and presence of outliers, given the amount of movement and possibility of occlusions.

The overall results indicate a similar performance across model variants both for the 2D and 3D-Stages ($\sim$ 3.73 pixels and 44.0 mm). This might indicate performance saturation for the task, given the amount of data and the noise present in the dataset.

The model is quite robust to corruption in either the inputs (Figure 4.11), being able to work with no depth or no image information, having to rely entirely on the remaining input feature. In these cases, the prediction confidence is lower, with also lower temporal consistency across frames. Although an improbable situation during

rehabilitation settings, it shows some of the potentials of using learning-based methods from a robustness standpoint.

Both 2D keypoint and connection heatmaps features, obtained from the intermediate objective, displayed the expected and interpretable Gaussian shape (Figure 4.7). This suggests that the model has correctly learned the features from the secondary regularization loss, since these are complementary to the keypoint detection objective.

The EfficientNet-lite0 feature extraction backbone displayed by far the lowest computation time in the CPU (Table 4.2), being faster than the ResNet18 baseline by 2.5 times and the ResNet50 by 6.4 times. Interestingly, the ResNet backbones offer much closer run-time performance when tested on the GPU. This might be explained by the fact that bigger, but simpler convolution operations are used, optimized to run on the GPU which processes large operations simultaneously. On the other hand, the EfficientNet backbone separates each convolution block into multiple small steps which are run sequentially introducing some overhead. When it comes to inference on the CPU, the great reduction in operations in the EfficientNet modules is greater than the overhead for the multiple sequential calls, yielding noticeably faster run-times, making it more adequate for the walker hardware. For applications targeting the GPU, the ResNet or its variants could be a better option.

The shallow refine module is capable of aggregating information to produce better results overall (Table 4.3), by also using connection cues from the connection heatmaps branch. These performance improvements come nonetheless at the cost of an increase around 45% in latency, which was considered acceptable since it was still within the performance requirements imposed.

The baseline lifting approach without depth information achieved similar results as the one with depth, both in relative and absolute spaces (Table 4.1). This means that depth is not required in this task, possibly since the poses have low variance, ambiguous poses are uncommon and the wrist positions are almost constant, while the depth information might not be too reliable since only a single noisy point is considered for each keypoint. This indicates that future pipelines based solely on RGB camera data would be possible, further decreasing the cost of hardware.

The 3D keypoints predicted by the model, despite showing similar trajectories to those generated by the Xsens ground truth, still displayed some constant positional offset and failure to capture the full range of motion on others. It might also predict incorrect limb lengths, that do not correspond to the subject's anthropometric data.

Although the leg and feet keypoints are detected on a different image reference frame from the torso, the lifting models are capable of internally finding a way to relate the information without the need of explicitly providing the extrinsic transformation between camera frames.

Furthermore, it was shown (Table 4.4) the lifting approach yields better results than the explicit projection method with residual correction. This could be explained by the fact that the lifting method always considers the 2D location information along

with the noisy depth values to produce the transformation. On the other hand, the projection method, although having a simplified task since the referential transformations are computed externally, in practice, is faced with frequent cases where one or more keypoints contain incorrect or no information due to bad projection (e.g., dead pixels, background pixel selection, different body thickness across subjects and body parts).

Unexpectedly, the complete temporal model performed worse when benchmarked against the single-frame counterparts (Table 4.7). Multiple reasons might have contributed to this. Namely, the use of a backbone pre-trained with longer temporal sequences, the harder optimization process given the more complex task of also relating temporal information, the use of a model with higher capacity given the same amount of data which might have resulted in some over-fitting. Nevertheless, these create temporally coherent results, without the need for any post-processing, preferable when dealing with sequential data. Unfortunately, the 3D convolutions added a significant amount of latency ($>$ 500 ms) especially in the CPU, making them impractical for the walker.

The model worked as expected on the smart walker (Figure 4.10b), running in real-time while displaying similar detection performance (qualitatively speaking) to the dataset test samples, when dealing with common walking situations used for training and complying with the hardware requirements. However, the performance degraded when presented with situations outside the training distribution, implying that more diverse data is needed to train a model capable of fully performing in real-world situations. Moreover, increased latency and response delay were reported, due to the asynchronous nature of the underlying ROS system in the equipment.

Some model ideas had to be changed or dropped entirely, as these were not correctly supported by the *.onnx* framework (and also most alternatives available), and thus not fit for the deployment scheme used. This included the use of: (i) SE attention blocks proposed by [127] for the 2D-Stage, tried initially with promising results in terms of performance with minimal effect on latency; (ii) EfficientNet (non-lite) versions, which also used the SE modules; (iii) SemGCN [125] modules for the 2D to 3D lifting.

4.6 Conclusion and future perspectives

This research presents a novel DL-based full-body pose estimation solution for the ASBGo smart walker. It is able to extract a compact body representation from two camera streams, which can be used for downstream tasks in patient monitoring and enable human-in-the-loop control strategies.

The proposed DL framework allows extracting a compact representation of the full human body, directly from inexpensive cameras and adaptable (given some training data) to multiple setup configurations. This is in opposition to competing smart walker solutions that rely on custom dedicated hardware, must be carefully tuned and are only

capable of monitoring a small set of metrics (e.g., gait analysis).

Limitations in the proposed research deal with the fact that the models are trained to produce confident predictions for data that fall inside the distribution used for training and will produce unreasonable predictions for data that falls outside. This is still an open issue in the DL field and thus, the need to fully benchmark deployed solutions prior to real-world usage is highly emphasized.

Moreover extensive hyper-parameter optimization was not performed, and better results could have been obtained with hyper-parameter search. However, given the computational resources and the absence of specific literature in the area, this research considered more appropriate to fairly compare different approaches, using reasonable configurations based on values commonly found in the DL literature.

The dataset also contains relatively low variability in terms of poses and their locations on the image frames, limiting its application to general human pose estimation problems. Data augmentation techniques were applied to the 2D-Stage to decrease visually and positional overfitting to the common keypoint locations on the image frames, while also collecting additional 13 k frames with irregular walking data which brought some improvements. Nevertheless, these were not sufficient to combat model overfitting to the overwhelming amount of common keypoint locations during walking. This made it unreasonable to explore more complex models, and fully train the 2D-Stage, as it would easily overfit. A possible solution to mitigate this problem would involve pretraining each stage or the full model on a general human pose estimation dataset [128] with similar data modalities (image, depth) and then fine-tune the last few layers on this dataset.

The use of solid models in the literature (e.g., OpenPose) for pose estimation could be a possibility to be extensively evaluated in future but always taking into account the computing power available. While the limited computing power may be perceived as a limitation, on the other hand, as highlighted in Sec. 2.5, it enables to distribute such advanced monitoring systems, without barriers due to the scarcity of economic resources and, therefore, the impossibility of acquiring more powerful hardware.

This research lays the groundwork for building an integrative framework for evaluating patients undergoing rehabilitation. The benefits of such applications are multiple: (i) they allow the clinician to build personalized rehabilitation protocols, (ii) they allow to adjust the rehabilitation program in real-time, (iii) they guarantee quantitative measures without the need of using intrusive sensors. Surely further work should be carried out to validate the proposed framework, however, promising results were obtained on healthy participants. Future improvements will deal with: the collection of more and variable data also from subjects with gait impairments. An extension to the dataset is thus planned, using the same acquisition setup, during real rehabilitation sessions. It will be used to further training the models and allow validation in clinical scenarios. Moreover prior knowledge on patients' kinematics may be provided to neural network as to have more coherent outcomes. While, from an implementation

point of view the combination of a single-frame 2D-Stage with a temporal 3D-Stage was considered as to allow temporally consistent 3D outputs with low computational time and no-response delay.

Chapter 5

End-to-end facial landmark detection to assess dysarthria evolution

Always quoting the contribution of Silvio Garattini: “There is little attention paid to the approximately 7000 rare diseases that span all specialties. These are only a few patients per disease who nevertheless have the same right to health as those with more common diseases”. This last research is the most recent among those presented in this thesis and is part of the wider project “Homely Care” aimed at innovating follow-up strategies for patients suffering from neurodegenerative disorders (e.g., ALS, spinal muscolar atrophy, stroke...). Similarly to all the previous works, this research was born and developed via a dialogue between a multidisciplinary team of clinicians and engineers who work in synergy to propose a system for monitoring the evolution of such complicated, disabling and unforgiving diseases.

The system is aimed at mapping neurodegenerative diseases progress via the monitoring of dysarthria evolution which is the set of speech disorder mainly induced by these rare diseases. “Homely Care” consists in a remotely-usable web application that patients download on their devices. The application allows them to take video-selfies while performing assessment tasks agreed with clinicians. The video selfies are sent to the cloud platform to be processed by DL algorithms, which return progression indices to the clinician viewable on a dedicated interface.

Currently, the assessment of neurodegenerative disease is carried out through direct observation of the patient by neurologists and speech therapists coupled with paper-and-pencil rating scales. Especially when dealing with such rare diseases, these qualitative, non-homogeneous and mostly paper-based evaluations suffer from a serious drawbacks: they do not allow rapid consultation and sharing of data and therefore undermine the possibilities of research in this field. In Homely Care, integrated smartphone cameras and microphones become new sensors that allow the clinician to be constantly informed about the patient’s health condition.

5.1 Background and motivation

Dysarthria is a neurological disorder caused by the generalized weakness and spasticity of the anatomical structures responsible for words and sounds production. Neurodegenerative diseases (e.g., ALS), inflammatory conditions (e.g., multiple sclerosis) and vascular pathologies (e.g., stroke) are the leading causes for dysarthria onset [25]. This disorder has different symptoms depending on the onset causes and the involved nervous-system structures. Dysarthric patients may suffer from: (i) changes in the strength, speed, amplitude, stability and tone of the voice, and (ii) impaired coordination of breathing, phonatory, resonatory and articulatory movements, with a pervasive and unrelenting impact on intelligibility, prosody and speech quality [129].

The ability to communicate as to keep and extend social contacts has a strong influence on the psychic balance of these patients and on their overall well-being. The functional evaluation of the evolution of dysarthria is relevant to readily identify the instant for prescribing compensation strategies to communicative disabilities and to monitor the evolution of the diseases for drawing up patient-specific care plans [130, 131]. With such a view, the assessment of oro-facial muscles impairments may significantly support clinicians to early identify functional changes in patient performance and accelerate the implementation of corrective and compensatory strategies [131].

Although its relevance is recognised in literature, the monitoring of oro-facial muscles mainly relies upon visual inspection by clinicians. This procedure, sometimes combined with the drafting of paper-and-pencil rating scales (e.g., Robertson Profile for Dysarthria) [132, 133], has the drawbacks of being qualitative, non-reproducible and highly influenced by the patients' emotional status at the time of examination. A possible solution to attenuate the issue of perspective evaluations has been proposed in [134]. The authors employ electromyographic sensors placed over the facial surface and inside the oral cavity. However, the use of such intrusive sensors makes the assessment unpleasant for patients and, additionally, the complex nature of the acquisition set-up mines its usage in the actual clinical practice.

To objectively and non-intrusively evaluate the status of patients with oro-facial impairments, the work in [135] proposes a DL methodology for assessing facial alignment from RGB videos of patients with ALS and stroke. For stimulating the research in the field, the authors in [135] further release their dataset, i.e., the Toronto NeuroFace dataset, which is the first annotated dataset in the field.

Inspired by the work in [135] and by applications of video-based facial-landmark detection in closer fields (e.g., for evaluating depression symptoms and assessing the presence of cerebral palsy [136]) this research present an end-to-end CNN pipeline for facial-landmark detection in patients with ALS and stroke using the Toronto NeuroFace dataset.

5.2 Methods

5.2.1 The Toronto NeuroFace dataset

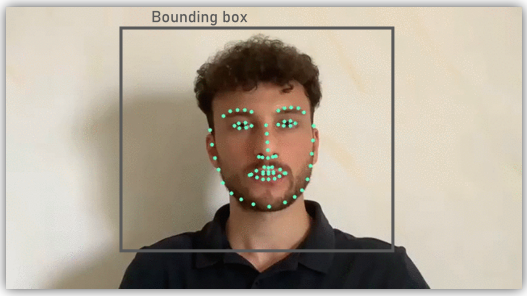


Figure 5.1: Sample image with the 68-facial landmarks (green dots) and the bounding box face annotation (black square).

The Toronto NeuroFace dataset in [135] contains RGB video recordings from 11 ALS patients (4 males, 7 females), 14 stroke patients (10 males, 4 females) and 11 healthy subjects (7 males, 4 females) that perform motor tasks (e.g., maximal mouth-opening, lips-stretching, lip-protrusion etc.). The manual annotation of 68-facial landmarks and face bounding box (Fig. 5.1) is performed for 3306 frames of which: 1015 frames healthy subjects, 920 frames for ALS patients and 1371 for stroke patients.

For the purpose of this work, the Toronto NeuroFace dataset was split in training, testing and validation set keeping frames from 32 subjects to train and validate the CNN and frames from 4 subjects (of which 2 with ALS and 2 with stroke) to test it.

5.2.2 Mask-RCNN for facial landmark detection

The pipeline for facial-landmark detection in people with ALS and stroke is shown in Fig. 5.2 and relies on Mask-RCNN [137], which was originally designed to predict the pose of the human body in two-dimensional space. Here, Mask-RCNN was modified to detect facial-landmarks.

Mask-RCNN has 3 main branches: classification, bounding-box regression and facial- landmarks position regression. Firstly, the RGB-input image is fed into a backbone-CNN. The backbone is a ResNet50 (i.e., ResNet with 50 convolutional layers) which acts as feature-pyramid network extractor (FPN). The FPN allows to retrieve feature maps which couple low-resolution, semantically strong features with high-resolution, semantically weak features. These output feature maps are fed into a two-stage architecture. This first stage consists in a CNN, named Region Proposal Network (RPN), which generates multiple Region of Interest (RoI). Then, each region proposal is sent to the RoI Align layer which extracts a small feature map from each

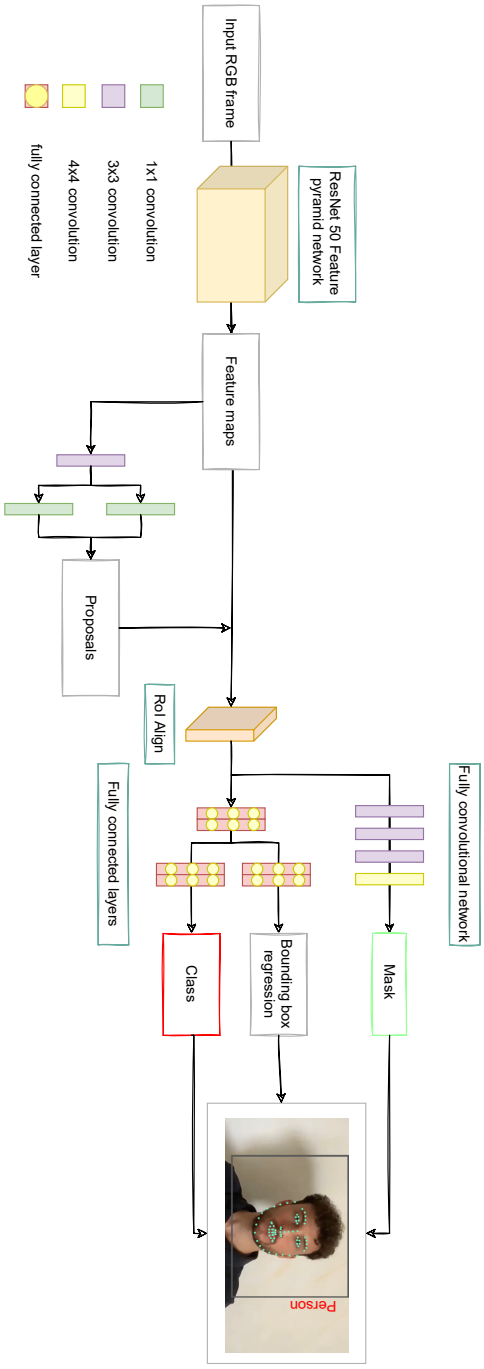


Figure 5.2: Mask-RCNN for facial-landmark detection.

RoIs. Warped features in output from the RoI Align are then fed into fully connected layers. These layers output the bounding-box coordinates and the label category (i.e., "face") with the relative prediction-confidence in output from soft-max layer of the classification branch. It is worth noting that this stage of bounding-box regression is crucial when deploying the pipeline in scenarios (e.g, hospital wards) where empty backgrounds may not be guaranteed. In parallel with the bounding box coordinate regression and class assignment, warped features are also fed into the mask branch which is a fully convolutional network properly adapted for landmarks detection task. Particularly this branch was modified for the task of interest by adding two subsequent strided (s=2)-transposed convolutions (with kernel sizes 3x3 and 4x4, respectively). The last convolutional layer outputs 68-one-hot binary masks, one for each landmark. The use of two additional transposed convolution layers allows the resolution of the output map to be extended, making it easier for the network to locate the 68-face landmarks.

For training, the fine-tuning technique was adopted. Thus the weights of the network was initialized with the weights resulting from pre-training on the 300 Faces In-the-Wild Challenge (300-W) dataset [138][1]. This is a public available dataset of faces captured in an outdoor and indoor environment, which shows great variability in terms of expressions, identities, facial positions and lighting levels. This modified Mask-RCNN fine-tuned on 300-W dataset was called 300W-N-FLMask.

5.3 Experimental Protocol

5.3.1 Training settings

The 300W-N-FLMask fine tuning was carried out in 28000 iterations. The initial learning rate was set to 0.001 with a learning rate decay of 0.5 every 15000 and 20000 iterations, respectively. The SGD was used as optimizer. The number of maximum subject-detection per image was limited to 1. The confidence threshold for the bounding box was set to 0.75. All these training settings come from an extensive grid-search to find the best combination of loss, optimizer, learning rate scheduling and iterations.

Online data augmentation was used to increase the size of the Toronto NeuroFace dataset. Random brightness (with brightness factor ranging from 0.8 to 1.2) and flipping (with a probability equal to 0.5) were considered. All the trainings were conducted using the Amazon Web Services cloud computing.

5.3.2 Ablation studies

The performance of 300W-N-FLMask was compared against the performance of its akin trained from scratch (N-FLMask) and the original N-MaskRCNN (i.e., without

[1] https://ibug.doc.ic.ac.uk/resources/300-W/

Table 5.1: Conducted ablation studies for validating the proposed architecture for facial landmark detection in patients suffering from neurodegenerative diseases.

	300W-N-FLMask	N-FLMask	300W-FLMask	N-Mask
Pretraining	300-W	x	x	x
Training	Toronto NeuroFace	Toronto NeuroFace	300-W	Toronto NeuroFace

Table 5.2: Results in terms of Normalized Mean Error (*NME*) for each of the 4 tested convolutional neural network (CNNs).

	300W-N-FLMask	N-FLMask	300W-FLMask	N-Mask
NME_{68}	**1.79**	2.70	3.88	13.55
NME_{chin}	**2.62**	4.81	4.39	15.31
$NME_{eyebrows}$	**0.02**	0.03	0.05	0.12
NME_{nose}	**1.55**	2.08	3.60	5.61
NME_{eyes}	**1.03**	0.94	3.04	5.23
NME_{mouth}	**1.49**	2.19	3.70	21.17

changing the number of convolutional layers in the branch for facial landmarks detection) trained from scratch. The performance of the proposed modified version of the Mask-RCNN (i.e., FLMask) was tested when trained on the 300-W dataset only (i.e., 300W-FLMask). An overview of the ablation studies in shown in Table 5.1. For each of the ablation studies the training settings are those described in Sec. 5.3.1.

5.3.3 Evaluation metrics

To assess the performance of the tested CNNs, the Normalized Mean Error (NME_k) was computed as follows:

$$NME_k = \left[\frac{1}{N_L} \sum_{i=1}^{N_L} \frac{\sqrt{(x_i - xp_i)^2 + (y_i - yp_i)^2}}{Diag_{bbox}} \right] \cdot 100$$

where k stands for the total number of images in the test set, $Diag_{bbox}$ is the length of the bounding box diagonal, while N_L identifies the number of landmarks, (x_i, y_i) are the ground-truth coordinates and (xp_i, yp_i) are the predicted landmark coordinates. The NME_k for the totality of 68 landmarks (NME_{68}), for the 17 chin landmarks (NME_{chin}), for the 10 eyebrow landmarks ($NME_{eyebrows}$), for the 9 nose landmarks (NME_{nose}), for the 12 eyes landmarks (NME_{eyes}) and for the 20 mouth landmarks (NME_{mouth}), were assessed.

5.4 Results

The results of the 4 tested CNNs are shown in the Table 5.2.

The performance of the N-FLMask was compared with the performance of the N-Mask to prove the effectiveness of the architectural variation. As visible from Tab. 5.2, the results of the N-FLMask was significantly better than the results of the N-Mask with a NME_{68} equal to 2.70 against NME_{68} equal to 13.55 achieved by the N-Mask.

To prove the effectiveness of the elected training protocol dealing with: (i) pre-training on 300-W Dataset and (ii) fine-tuning on the Toronto NeuroFace dataset, the performance of the 300W-N-FLMask was compared with that of the N-FLMask and 300W-FLMask. As showed in Tab. 5.2, the 300W-N-FLMask achieved the highest performance with a NME_{68} equal to 1.79 against NME_{68} equal to 2.70 and 3.88 of the N-FLMask and 300W-FLMask, respectively.

5.5 Discussion

Evaluating the evolution of dysarthria by monitoring the impact that this pathology has on the muscles of the oro-facial district is fundamental to assessing the evolution of the neurodegenerative disorders as ALS or stroke. However, despite its importance, this assessment is carried out by direct observation of the patient coupled with rating scales. This procedure, besides being qualitative e discontinuous, does not allow to perceive fine changes in patients' performance.

To solve issues caused by perspective assessments electromyographic sensors are proposed in literature for evaluating the oro-facial muscles in patients suffering from dysarthria. However, this analysis, based on sensors placed on the facial surface and inside the oral cavity, is too invasive for the patient and can only be carried out in controlled environments.

To overcome possible state-of-art limitations and with the aim of making quantitative, simple and non-invasive the assessments of patients suffering from dysarthria, the proposed research presents a DL methodology capable of regressing the position of facial landmarks in subject with neurodegenerative diseases.

Comparing the results of the 300W-N-FLMask architecture with the performance of N-FLMask and 300W-FLMask it emerged that the choice of pretraining the architecture on the 300-W dataset and then applying fine-tuning on the Toronto NeuroFace dataset was crucial. In fact, the pre-training on the wider dataset granted the network a higher power of generalisation while the fine-tuning allowed to refine the CNN ability in regressing facial landmarks from pathological subjects.

The original version of the Mask-RCNN (i.e., N-Mask) got the largest error (i.e., the lowest performance). This suggests that varying the branch for landmarks-position regression allowed to recover an higher level of details. It is worth noting that the chin and mouth were the most challenging to detect, and this may depend on the major impact that the progressive-disease has on the muscles of the oral district [135]. In this case, the least flawed of the 4 architectures was the 300W-N-FLMask. This may be due to the fact that the 300-W dataset, on which the network was pre-trained, had

frames from people grimacing and therefore with oro-facial skewness.

5.6 Conclusion and future perspectives

Dysarthric patients need to clearly report to clinicians the elements useful to establish their health condition. They must be able to communicate as to keep and extend social contacts. This aspect has a strong impact on patient's psychological balance and, consequently, on his/her overall well-being.

Monitoring the evolution of dysarthria through a non-intrusive assessment of the oro-facial musculature is of primary importance to allow the clinician to implement preventive strategies to compensate for communicative disability. This research proposes, for among the first time in literature, a CNN-based methodology to detect facial-landmarks position in RGB images of ALS and stroke patients. As highlighted in [135], from the position of facial landmarks it will be possible to evaluate the impact that these neurodegenerative diseases have on the muscles involved in vital functions such as speech articulation and breathing as to provide decision support to clinicians. In this regard, of particular clinical interest will be the assessment of: lip protrusion, lip stretching, maximum mouth opening, facial and mouth symmetry and how these indexes evolve in time.

This research represents a small part of a much wider project aimed at proposing the first remote digital assessment tool to support clinicians in mapping the evolution of neurodegenerative diseases. The tool will include a web application that allows the patient to take video-selfies while performing Robertson's dysarthria profile inspired tasks. The audio-video data stream is processed via DL algorithms which extract indexes of progression for enabling the clinicians to soon identifying changes in patients' performance and readily prescribing compensation strategies to communicative disabilities.

Especially when dealing with rare diseases, having quantitative and easily accessible data is relevant both (i) for establishing treatment plans specific to the needs of individual patients and offering them the best possible care, (ii) for gaining knowledge about such a disabling disease. Moreover, from patients' side, travelling to the hospital is sometimes tiring, decreases their performance during evaluations phases and has a real cost. Systems such as the one described in this research are crucial as they enable patients to carry out assessment at home, with familiar devices (such as smartphone and pc) and in a familiar environment while always being in virtual contact with their trusted clinician.

Chapter 6

Conclusive remarks

6.1 Conclusion

This thesis chronicled a three-year journey. Three years spent inside a NICU, a center specialized in the treatment of children with autism and neurology departments. Three years of interdisciplinary dialogue between clinicians and engineers to imagine new models of monitoring fragile patients who need special care.

The journey began in the NICU of the "G. Salesi" Hospital in Ancona. A neonatologist in front of many cribs housing preterm infants so small they fit within a hand. The neonatologist is in front of the first crib, observes the movement of the infant and notes the observation on the patient's health record, moves on, to the second crib, to the third, to the fourth and repeats the manual annotation procedure to the last crib.

The journey continues inside a center specialised in ABA therapy for treating children with autism. There is an ABA operator sitting holding a pen and writing on a pad of paper. He writes without looking on the paper as the gaze is fixed on a child in front of him. We change room. The operator picks up the paper and pen and follows the child. He puts down the pad to help the child wash his hands but picks it up again as soon as the child finished the action.

The last stages of the journey take place within the rehabilitation medicine and neurology departments. The situation is similar to those previously described. There are a clinician and a patient. Facing each other. The clinician asks the patient to perform evaluative tasks and notes what he observes on the patient's health record.

Medicine today is still heavily based on direct observation of the patients combined with the compilation of rating scales (e.g., [133]). This assessment procedure, besides discontinuous, strongly depends on the experience of the examiner.

To overcome the limitations posed by qualitative and sporadic assessments, this thesis showed innovative monitoring systems based on the analysis of video-recordings acquired by RGB-D cameras. The described methodologies are aimed at supporting: (i) the neonatologist in monitoring the movement of preterm infants, (ii) the ABA operators in assessing when the children with autism wash their hands independently, (iii) the rehabilitation expert in evaluating the posture of a person during rehabilitation sessions, (iv) the neurologist and speech therapist in mapping the evolution of

neurodegenerative disease.

The national health care system is an absolutely irreplaceable and valuable asset for public health and the health of each of us. It showed its weaknesses during the COVID-19 pandemic due primarily to clinical staff shortages. The monitoring systems presented in this thesis was born and developed from this compelling need of expanding clinicians' operational capacity and will continue to be refined to further meet patients' needs. Patients' voice, in fact, will become louder, and more conscious and will make the clinician-patient relationship active and bidirectional. Patients, or their caregivers, will be the primary collectors of data, and all of this data will be gathered in unique, secure, easily accessible and dialoguing systems.

The benefits resulting from increased digitization of healthcare are manifold: clinicians will dispose a complete and structured collection of data on the patient's history and will be able to compare similar clinical pictures in order to identify the most appropriate treatment for the patient. Patients, actively involved in their care plan, will begin to consider health as a system and not as a pillar and that prevention is better than cure. They will begin to trust in healthy habits and healthy relationship with the surrounding environment as essential elements to preserve their wellbeing [18].

The need for health is much greater than the answers that are given from our healthcare system. With the current paradigms, the best standard of care cannot be guaranteed to those who need it. In the vision that this thesis seeks to convey, digital health should support the healthcare system by making it more efficient, releasing energies that will allow clinicians to engage with patients, who are increasingly participatory and conscientious, and treat them better and to the best of their abilities [5].

6.2 Impact

During these years of pandemia, above all, we have seen our healthcare system in trouble, overwhelmed by too many calls for help. We have seen clinicians and nurses working prohibitively long shifts to cope with staff shortages and we have seen patients in great need of treatment but too afraid to enter hospital because of the risk of being infected by the virus.

Health is essential, concerns everyone without distinction and must be guaranteed to all. The events following the pandemic have forced us to think both a new concept of "normality" and the future of the entire health system. A future in which new technologies must be increasingly present and must both broaden clinicians capacities and ensure continuity of care.

We can no longer afford to lose data on a rare disease. It is precisely because of its rarity that any data must be fundamental to the progress of research. We cannot allow a clinician to spend so many hours observing a patient because if resources are scarce, they must be spent treating and talking with patients, not merely observing them.

This thesis was born from the desire to support clinicians during the actual clinical

practice and to enable that patients and their families always feel cared for. All the proposed monitoring systems, besides being non-intrusive, are developed to ensure quantitative measurements useful in clinical practice and are designed to be easily portable also in the home environment. Alongside the DL algorithms, the real protagonist of the thesis is the multimedia data that becomes a source of new knowledge for a clinician. A kind of data easily available as our society is incredibly good at generating it in large quantities and giving it away just as easily.

Of course, there is still a lot of work to largely distribute these systems as to trigger healthcare digital transformation. For this large-scale innovations, for sure, there is no on-off switch or precise moment when healthcare goes from being analogue to digital. The healthcare transformation needs: dialogue, clinicians' and patients' preparation for distributing innovative solutions without age and culture barriers and the entire ecosystem adaptation which is still too much based on cumbersome mechanisms.

Once the digital transformation of the health ecosystem has taken place, we should no longer have to imagine how defined a disease picture can be thanks to the consultation and filtering of universally collected homogeneous new data or how much knowledge could be generated by accessing thousands of similar situations in real time as this will be part of a new everyday life. A reality in which the patient generates health-related data everywhere and the clinician receives it in a structured way. A patient increasingly involved in his/her own care plan and increasingly responsible for his/her own health. A reality in which the clinician no longer has to search through piles of paper for the results of an evaluation conducted years earlier but can dialogue more with their patients. A reality in which research becomes a bridge to strengthen the empathic relationship between clinician and patient.

6.3 Future perspectives

At the center of every significant change in our lives today is a technology of some kind. Thanks to technology, everything we do is always in the dimension of becoming: everything is becoming something else. This perpetual change is the central pillar of the modern world. Especially when we talk about health and healthcare in general, which are issues that affect the entire world, changes are not always well received. The very first impulse when facing with extreme digital technologies (like deep learning) is always to reject them, stop them, ban them or at least make them difficult to use [3]. However this denial is temporary, we begin to realize that by working with technology, by actively participating in the innovation process we can get the best out of what technology can offer [5].

We are dealing with a total paradigm shift, a revolution that will be prominently technological but not completely. The real and more complex transformation will be cultural, mental. So if the trigger of this revolution is digital, the node will be human because humans and nor algorithms are the protagonists.

Bibliography

[1] S. Zhou, D. Nie, E. Adeli, J. Yin, J. Lian, and D. Shen, "High-resolution encoder–decoder networks for low-contrast medical image segmentation," *IEEE Transactions on Image Processing*, vol. 29, pp. 461–475, 2020.

[2] S.-Y. Lo, H.-M. Hang, S.-W. Chan, and J.-J. Lin, "Efficient dense modules of asymmetric convolution for real-time semantic segmentation," in *Proceedings of the ACM Multimedia Asia*, 2019, pp. 1–6.

[3] K. Kelly, *L'inevitabile*. Il Saggiatore, 2017.

[4] N. Bostrom and E. Yudkowsky, "The ethics of artificial intelligence," *The Cambridge Handbook of Artificial Intelligence*, vol. 1, pp. 316–334, 2014.

[5] R. Ascione, "Il futuro della salute: come la tecnologia digitale sta rivoluzionando la medicina (e la nostra vita)," *Il futuro della salute*, pp. 1–270, 2018.

[6] T. Davenport and R. Kalakota, "The potential for artificial intelligence in healthcare," *Future healthcare journal*, vol. 6, no. 2, p. 94, 2019.

[7] K. Ostherr, "Artificial intelligence and medical humanities," *Journal of Medical Humanities*, pp. 1–22, 2020.

[8] S. Aminololama-Shakeri and J. E. López, "The doctor-patient relationship with artificial intelligence," *American Journal of Roentgenology*, vol. 212, no. 2, pp. 308–310, 2019.

[9] E. Topol, *Deep medicine: how artificial intelligence can make healthcare human again*. Hachette UK, 2019.

[10] F. Piccialli, V. Di Somma, F. Giampaolo, S. Cuomo, and G. Fortino, "A survey on deep learning in medicine: Why, how and when?" *Information Fusion*, vol. 66, pp. 111–137, 2021.

[11] D. L. Sackett, "Evidence-based medicine," in *Seminars in Perinatology*, vol. 21, no. 1. Elsevier, 1997, pp. 3–5.

[12] T. Hulsen, S. S. Jamuar, A. R. Moody, J. H. Karnes, O. Varga, S. Hedensted, R. Spreafico, D. A. Hafler, and E. F. McKinney, "From big data to precision medicine," *Frontiers in Medicine*, vol. 6, p. 34, 2019.

[13] J. C. Denny and F. S. Collins, "Precision medicine in 2030—seven ways to transform healthcare," *Cell*, vol. 184, no. 6, pp. 1415–1419, 2021.

[14] G. Hinton, "Deep learning—a technology with the potential to transform health care," *Jama*, vol. 320, no. 11, pp. 1101–1102, 2018.

[15] L. C. Yan, B. Yoshua, and H. Geoffrey, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[16] J. T. Kim, "Application of machine and deep learning algorithms in intelligent clinical decision support systems in healthcare," *Journal of Health & Medical Informatics*, vol. 9, no. 05, 2018.

[17] A. Salam, "Internet of things for sustainable human health," in *Internet of Things for Sustainable Community Development*. Springer, 2020, pp. 217–242.

[18] I. Capua, "Salute circolare," *Una rivoluzione necessaria. Editore Egea-Bocconi, Milano*, 2019.

[19] M. Ye, J. Shen, G. Lin, T. Xiang, L. Shao, and S. C. Hoi, "Deep learning for person re-identification: A survey and outlook," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021.

[20] J. Xu, F. Huang, X. Zhang, S. Wang, C. Li, Z. Li, and Y. He, "Sentiment analysis of social images via hierarchical deep fusion of content and links," *Applied Soft Computing*, vol. 80, pp. 387–399, 2019.

[21] L. Migliorelli, S. Moccia, I. Avellino, M. C. Fiorentino, and E. Frontoni, "Mydi application: Towards automatic activity annotation of young patients with type 1 diabetes," in *2019 IEEE 23rd International Symposium on Consumer Technologies (ISCT)*. IEEE, 2019, pp. 220–224.

[22] Q. Xie, K. Faust, R. Van Ommeren, A. Sheikh, U. Djuric, and P. Diamandis, "Deep learning for image analysis: Personalizing medicine closer to the point of care," *Critical Reviews in Clinical Laboratory Sciences*, vol. 56, no. 1, pp. 61–73, 2019.

[23] I. Zuzarte, P. Indic, D. Sternad, and D. Paydarfar, "Quantifying movement in preterm infants using photoplethysmography," *Annals of Biomedical Engineering*, vol. 47, no. 2, pp. 646–658, 2019.

[24] S. Artoni, L. Bastiani, M. C. Buzzi, M. Buzzi, O. Curzio, S. Pelagatti, and C. Senette, "Technology-enhanced aba intervention in children with autism: a pilot study," *Universal Access in the Information Society*, vol. 17, no. 1, pp. 191–210, 2018.

[25] B. Tomik and R. J. Guiloff, "Dysarthria in amyotrophic lateral sclerosis: A review," *Amyotrophic Lateral Sclerosis*, vol. 11, no. 1-2, pp. 4–15, 2010.

[26] L. Migliorelli, S. Moccia, R. Pietrini, V. P. Carnielli, and E. Frontoni, "The babypose dataset," *Data in Brief*, vol. 33, p. 106329, 2020.

[27] J. Tucker and W. McGuire, "Epidemiology of preterm birth," *Bmj*, vol. 329, no. 7467, pp. 675–678, 2004.

[28] J.-A. Quinn, F. M. Munoz, B. Gonik, L. Frau, C. Cutland, T. Mallett-Moore, A. Kissou, F. Wittke, M. Das, T. Nunes *et al.*, "Preterm birth: Case definition & guidelines for data collection, analysis, and presentation of immunisation safety data," *Vaccine*, vol. 34, no. 49, pp. 6047–6056, 2016.

[29] A. Polito, S. Piga, P. E. Cogo, C. Corchia, V. Carnielli, M. Da Frè, D. Di Lallo, I. Favia, L. Gagliardi, F. Macagno *et al.*, "Increased morbidity and mortality in very preterm/VLBW infants with congenital heart disease," *Intensive Care Medicine*, vol. 39, no. 6, pp. 1104–1112, 2013.

[30] R. M. Ward and J. C. Beachy, "Neonatal complications following preterm birth," *BJOG: An International Journal of Obstetrics & Gynaecology*, vol. 110, pp. 8–16, 2003.

[31] C. Einspieler, A. F. Bos, M. E. Libertus, and P. B. Marschik, "The general movement assessment helps us to identify preterm infants at risk for cognitive dysfunction," *Frontiers in Psychology*, vol. 7, p. 406, 2016.

[32] I. Bernhardt, M. Marbacher, R. Hilfiker, and L. Radlinger, "Inter-and intra-observer agreement of Prechtl's method on the qualitative assessment of general movements in preterm, term and young infants," *Early Human Development*, vol. 87, no. 9, pp. 633–639, 2011.

[33] F. Ferrari, C. Einspieler, H. Prechtl, A. Bos, and G. Cioni, *Prechtl's method on the qualitative assessment of general movements in preterm, term and young infants*. Mac Keith Press, 2004.

[34] T. Moore, S. Johnson, S. Haider, E. Hennessy, and N. Marlow, "Relationship between test scores using the second and third editions of the Bayley Scales in extremely preterm children," *The Journal of Pediatrics*, vol. 160, no. 4, pp. 553–558, 2012.

[35] D. G. Sweet, V. Carnielli, G. Greisen, M. Hallman, E. Ozek, R. Plavka, O. D. Saugstad, U. Simeoni, C. P. Speer, M. Vento *et al.*, "European consensus guidelines on the management of respiratory distress syndrome-2016 update," *Neonatology*, vol. 111, no. 2, pp. 107–125, 2017.

[36] I. Trujillo-Priego, C. Lane, D. Vanderbilt, W. Deng, G. Loeb, J. Shida, and B. Smith, "Development of a wearable sensor algorithm to detect the quantity and kinematic characteristics of infant arm movement bouts produced across a full day in the natural environment," *Technologies*, vol. 5, no. 3, p. 39, 2017.

[37] B. Smith, I. Trujillo-Priego, C. Lane, J. Finley, and F. Horak, "Daily quantity of infant leg movement: wearable sensor algorithm and relationship to walking onset," *Sensors*, vol. 15, no. 8, pp. 19 006–19 020, 2015.

[38] M. Airaksinen, O. Räsänen, E. Ilén, T. Häyrinen, A. Kivi, V. Marchi, A. Gallen, S. Blom, A. Varhe, N. Kaartinen *et al.*, "Automatic posture and movement tracking of infants with wearable movement sensors," *Scientific reports*, vol. 10, no. 1, pp. 1–13, 2020.

[39] C. Jiang, C. J. Lane, E. Perkins, D. Schiesel, and B. A. Smith, "Determining if wearable sensors affect infant leg movement frequency," *Developmental Neurorehabilitation*, vol. 21, no. 2, pp. 133–136, 2018.

[40] A. Cenci, D. Liciotti, E. Frontoni, A. Mancini, and P. Zingaretti, "Non-contact monitoring of preterm infants using RGB-D camera," in *International Design Engineering Technical Conferences and Computers and Information in Engineering Conference*. American Society of Mechanical Engineers, 2015, pp. V009T07A003–V009T07A003.

[41] S. Orlandi, K. Raghuram, C. R. Smith, D. Mansueto, P. Church, V. Shah, M. Luther, and T. Chau, "Detection of atypical and typical infant movements using computer-based video analysis," in *Annual International Conference of the IEEE Engineering in Medicine and Biology Society*. IEEE, 2018, pp. 3598–3601.

[42] T. Tsuji, S. Nakashima, H. Hayashi, Z. Soh, A. Furui, T. Shibanoki, K. Shima, and K. Shimatani, "Markerless measurement and evaluation of general movements in infants," *Scientific Reports*, vol. 10, no. 1, pp. 1–13, 2020.

[43] D. Freymond, Y. Schutz, J. Decombaz, J.-L. Micheli, and E. Jéquier, "Energy balance, physical activity, and thermogenic effect of feeding in premature infants," *Pediatric Research*, vol. 20, no. 7, p. 638, 1986.

[44] V. Marchi, A. Hakala, A. Knight, F. D'Acunto, M. L. Scattoni, A. Guzzetta, and S. Vanhatalo, "Automated pose estimation captures key aspects of general movements at eight to 17 weeks from conventional videos," *Acta Paediatrica*, vol. 108, no. 10, pp. 1817–1824, 2019.

[45] E. A. Ihlen, R. Støen, L. Boswell, R.-A. de Regnier, T. Fjørtoft, D. Gaebler-Spira, C. Labori, M. C. Loennecken, M. E. Msall, U. I. Möinichen *et al.*,

"Machine learning of infant spontaneous movements for the early prediction of cerebral palsy: A multi-site cohort study," *Journal of Clinical Medicine*, vol. 9, no. 1, p. 5, 2020.

[46] K. D. McCay, E. S. Ho, H. P. Shum, G. Fehringer, C. Marcroft, and N. D. Embleton, "Abnormal infant movements classification with deep learning on pose-based features," *IEEE Access*, vol. 8, pp. 51 582–51 592, 2020.

[47] K. Raghuram, S. Orlandi, P. Church, T. Chau, E. Uleryk, P. Pechlivanoglou, and V. Shah, "Automated movement recognition to predict motor impairment in high-risk infants: a systematic review of diagnostic test accuracy and meta-analysis," *Developmental Medicine & Child Neurology*, vol. 63, pp. 637–648, 2021.

[48] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh, "Realtime multi-person 2D pose estimation using part affinity fields," in *Conference on Computer Vision and Pattern Recognition*, 2017, pp. 7291–7299.

[49] A. Newell, K. Yang, and J. Deng, "Stacked hourglass networks for human pose estimation," in *European Conference on Computer Vision*. Springer, 2016, pp. 483–499.

[50] B. Fallang, O. D. Saugstad, J. Grøgaard, and M. Hadders-Algra, "Kinematic quality of reaching movements in preterm infants," *Pediatric Research*, vol. 53, no. 5, p. 836, 2003.

[51] C. B. Heriza, "Comparison of leg movements in preterm infants at term with healthy full-term infants," *Physical Therapy*, vol. 68, no. 11, pp. 1687–1693, 1988.

[52] T. H. Kakebeeke, K. von Siebenthal, and R. H. Largo, "Differences in movement quality at term among preterm and term infants," *Neonatology*, vol. 71, no. 6, pp. 367–378, 1997.

[53] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei, "Large-scale video classification with convolutional neural networks," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 1725–1732.

[54] H. Rahmati, R. Dragon, O. M. Aamo, L. Adde, Ø. Stavdahl, and L. Van Gool, "Weakly supervised motion segmentation with particle matching," *Computer Vision and Image Understanding*, vol. 140, pp. 30–42, 2015.

[55] R. Hou, C. Chen, and M. Shah, "An end-to-end 3D convolutional neural network for action detection and segmentation in videos," *arXiv preprint arXiv:1712.01111*, 2017.

[56] E. Colleoni, S. Moccia, X. Du, E. De Momi, and D. Stoyanov, "Deep learning based robotic tool detection and articulation estimation with spatio-temporal layers," *IEEE Robotics and Automation Letters*, vol. 4, no. 3, pp. 2714–2721, 2019.

[57] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2015, pp. 234–241.

[58] A. Bulat and G. Tzimiropoulos, "Human pose estimation via convolutional part heatmap regression," in *European Conference on Computer Vision*. Springer, 2016, pp. 717–732.

[59] J. Hosang, R. Benenson, and B. Schiele, "Learning non-maximum suppression," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 4507–4515.

[60] S.-E. Wei, V. Ramakrishna, T. Kanade, and Y. Sheikh, "Convolutional pose machines," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 4724–4732.

[61] X. Du, T. Kurmann, P.-L. Chang, M. Allan, S. Ourselin, R. Sznitman, J. D. Kelly, and D. Stoyanov, "Articulated multi-instrument 2-D pose estimation using fully convolutional networks," *IEEE Transactions on Medical Imaging*, vol. 37, no. 5, pp. 1276–1287, 2018.

[62] V. Penza, X. Du, D. Stoyanov, A. Forgione, L. S. Mattos, and E. De Momi, "Long term safety area tracking (LT-SAT) with online failure detection and recovery for robotic minimally invasive surgery," *Medical Image Analysis*, vol. 45, pp. 13–23, 2018.

[63] S. Moccia, S. J. Wirkert, H. Kenngott, A. S. Vemuri, M. Apitz, B. Mayer, E. De Momi, L. S. Mattos, and L. Maier-Hein, "Uncertainty-aware organ classification for surgical data science applications in laparoscopy," *IEEE Transactions on Biomedical Engineering*, vol. 65, no. 11, pp. 2649–2659, 2018.

[64] R. Azad, M. Asadi-Aghbolaghi, M. Fathy, and S. Escalera, "Attention deeplabv3+: Multi-level context attention mechanism for skin lesion segmentation," in *European Conference on Computer Vision*. Springer, 2020, pp. 251–266.

[65] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 4700–4708.

[66] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *European Conference on Computer Vision*, 2018, pp. 801–818.

[67] Y. Pang, Y. Li, J. Shen, and L. Shao, "Towards bridging semantic gap to improve semantic segmentation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 4230–4239.

[68] H.-K. Kim, K.-Y. Yoo, J. H. Park, and H.-Y. Jung, "Asymmetric encoder-decoder structured fcn based lidar to color image generation," *Sensors*, vol. 19, no. 21, p. 4818, 2019.

[69] J. Wang, H. Xiong, H. Wang, and X. Nian, "Adscnet: asymmetric depthwise separable convolution for semantic segmentation in real-time," *Applied Intelligence*, vol. 50, no. 4, pp. 1045–1056, 2020.

[70] C. Zhang, S. Bengio, M. Hardt, B. Recht, and O. Vinyals, "Understanding deep learning (still) requires rethinking generalization," vol. 64, no. 3, p. 107–115, 2021.

[71] H. Yang, C. Yuan, B. Li, Y. Du, J. Xing, W. Hu, and S. J. Maybank, "Asymmetric 3d convolutional neural networks for action recognition," *Pattern Recognition*, vol. 85, pp. 1–12, 2019.

[72] M. Mormina, "Science, technology and innovation as social goods for development: rethinking research capacity building from sen's capabilities approach," *Science and Engineering Ethics*, vol. 25, no. 3, pp. 671–692, 2019.

[73] J. Chen and X. Ran, "Deep learning with edge computing: A review." *Proceedings of the IEEE*, vol. 107, no. 8, pp. 1655–1674, 2019.

[74] S. Cass, "Nvidia makes it easy to embed ai: The jetson nano packs a lot of machine-learning power into diy projects-[hands on]," *IEEE Spectrum*, vol. 57, no. 7, pp. 14–16, 2020.

[75] E. Strubell, A. Ganesh, and A. McCallum, "Energy and policy considerations for modern deep learning research," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 09, 2020, pp. 13 693–13 696.

[76] M. Rashid, M. A. Khan, M. Alhaisoni, S.-H. Wang, S. R. Naqvi, A. Rehman, and T. Saba, "A sustainable deep learning framework for object recognition using multi-layers deep features fusion and selection," *Sustainability*, vol. 12, no. 12, p. 5037, 2020.

[77] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, "Mobilenets: Efficient convolutional neural networks for mobile vision applications," *arXiv preprint arXiv:1704.04861*, 2017.

[78] X. Ran, H. Chen, X. Zhu, Z. Liu, and J. Chen, "Deepdecision: A mobile deep learning framework for edge video analytics," in *IEEE Conference on Computer Communications*. IEEE, 2018, pp. 1421–1429.

[79] F. Wang, M. Zhang, X. Wang, X. Ma, and J. Liu, "Deep learning for edge computing applications: A state-of-the-art survey," *IEEE Access*, vol. 8, pp. 58 322–58 336, 2020.

[80] A. van Wynsberghe, "Sustainable ai: Ai for sustainability and the sustainability of ai," *AI and Ethics*, pp. 1–6, 2021.

[81] S. Moccia, L. Migliorelli, R. Pietrini, and E. Frontoni, "Preterm infants' limb-pose estimation from depth images using convolutional neural networks," in *2019 IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology*, 2019, pp. 1–7.

[82] C. Lord, M. Elsabbagh, G. Baird, and J. Veenstra-Vanderweele, "Autism spectrum disorder," *The Lancet*, vol. 392, no. 10146, pp. 508–520, 2018.

[83] L. A. Livingston, E. Colvert, S. R. S. Team, P. Bolton, and F. Happé, "Good social skills despite poor theory of mind: exploring compensation in autism spectrum disorder," *Journal of Child Psychology and Psychiatry*, vol. 60, no. 1, pp. 102–110, 2019.

[84] G. R. Mayer and B. Sulzer-Azaroff, *Applying behavior-analysis procedures with children and youth*. Holt, Rinehart and Winston, 1977.

[85] D. Granpeesheh, J. Tarbox, and D. R. Dixon, "Applied behavior analytic interventions for children with autism: a description and review of treatment research," *Annals of Clinical Psychiatry*, vol. 21, no. 3, pp. 162–173, 2009.

[86] J. Xie, L. Wang, P. Webster, Y. Yao, J. Sun, S. Wang, and H. Zhou, "A two-stream end-to-end deep learning network for recognizing atypical visual attention in autism spectrum disorder," *arXiv preprint arXiv:1911.11393*, 2019.

[87] A. Zunino, P. Morerio, A. Cavallo, C. Ansuini, J. Podda, F. Battaglia, E. Veneselli, C. Becchio, and V. Murino, "Video gesture analysis for autism spectrum disorder detection," in *2018 24th International Conference on Pattern Recognition (ICPR)*. IEEE, 2018, pp. 3421–3426.

[88] C. P. Johnson, S. M. Myers *et al.*, "Identification and evaluation of children with autism spectrum disorders," *Pediatrics*, vol. 120, no. 5, pp. 1183–1215, 2007.

[89] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[90] Y.-X. Wang, D. Ramanan, and M. Hebert, "Growing a brain: Fine-tuning by increasing model capacity," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 2471–2480.

[91] J. L. Cook, S.-J. Blakemore, and C. Press, "Atypical basic movement kinematics in autism spectrum conditions," *Brain*, vol. 136, no. 9, pp. 2816–2824, 2013.

[92] E. Frontoni, A. Mancini, M. Baldi, M. Paolanti, S. Moccia, P. Zingaretti, V. Landro, and P. Misericordia, "Sharing health data among general practitioners: The nu. sa. project," *International Journal of Medical Informatics*, vol. 129, pp. 267–274, 2019.

[93] WHO, *World Report on Disability*. World Health Organization, 2011.

[94] T. Mikolajczyk, I. Ciobanu, D. I. Badea, A. Iliescu, S. Pizzamiglio, T. Schauer, T. Seel, P. L. Seiciu, D. L. Turner, and M. Berteanu, "Advanced technology for gait rehabilitation: An overview," *Advances in Mechanical Engineering*, vol. 10, no. 7, pp. 1–19, 2018.

[95] J. Jonsdottir and M. Ferrarin, *Gait Disorders in Persons After Stroke*. Cham: Springer International Publishing, 2017, pp. 1–11.

[96] W. Johnson, O. Onuma, M. Owolabi, and S. Sachdev, "Stroke: A global response is needed," *Bulletin of the World Health Organization*, vol. 94, no. 9, pp. 634A–635A, 2016.

[97] R. Moreira, J. Alves, A. Matias, and C. P. Santos, *Smart and Assistive Walker – ASBGo: Rehabilitation Robotics: A Smart– Walker to Assist Ataxic Patients*. Springer Nature Switzerland AG, 2019, pp. 37–68.

[98] M. M. Martins, C. P. Santos, A. Frizera-Neto, and R. Ceres, "Assistive mobility devices focusing on smart walkers: Classification and review," *Robotics and Autonomous Systems*, vol. 60, no. 4, pp. 548 – 562, 2012.

[99] Ł. Kidziński, B. Yang, J. Hicks, A. Rajagopal, S. Delp, and M. Schwartz, "Deep neural networks enable quantitative movement analysis using single-camera videos," *Nature Communications*, vol. 11, no. 1, 2020.

[100] Y. Chen, Y. Tian, and M. He, "Monocular human pose estimation: A survey of deep learning-based methods," *Computer Vision and Image Understanding*, vol. 192, p. 102897, 2020.

[101] A. F. Neto, A. Elias, C. Cifuentes, C. Rodriguez, T. Bastos, and R. Carelli, "Smart walkers: Advanced robotic human walking-aid systems," in *Intelligent Assistive Robots*. Springer, 2015, pp. 103–131.

[102] J. Paulo, P. Peixoto, and U. J. Nunes, "Isr-aiwalker: Robotic walker for intuitive and safe mobility assistance and gait analysis," *IEEE Transactions on Human-Machine Systems*, vol. 47, no. 6, pp. 1110–1122, 2017.

[103] A. Frizera-Neto, R. Ceres, E. Rocon, and J. L. Pons, "Empowering and assisting natural human mobility: The simbiosis walker," *International Journal of Advanced Robotic Systems*, vol. 8, no. 3, p. 29, 2011.

[104] S. D. Sierra M, M. Garzón, M. Munera, C. A. Cifuentes *et al.*, "Human-robot-environment interaction interface for smart walker assisted gait: Agora walker," *Sensors*, vol. 19, no. 13, p. 2897, 2019.

[105] W.-H. Mou, M.-F. Chang, C.-K. Liao, Y.-H. Hsu, S.-H. Tseng, and L.-C. Fu, "Context-aware assisted interactive robotic walker for parkinson's disease patients," in *2012 IEEE/RSJ International Conference on Intelligent Robots and Systems*. IEEE, 2012, pp. 329–334.

[106] D. Mehta, S. Sridhar, O. Sotnychenko, H. Rhodin, M. Shafiei, H.-P. Seidel, W. Xu, D. Casas, and C. Theobalt, "Vnect: Real-time 3d human pose estimation with a single rgb camera," *ACM Transactions on Graphics (TOG)*, vol. 36, no. 4, pp. 1–14, 2017.

[107] D. Groos, H. Ramampiaro, and E. A. Ihlen, "Efficientpose: Scalable single-person pose estimation," *Applied Intelligence*, 2020.

[108] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh, "Realtime multi-person 2d pose estimation using part affinity fields," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 7291–7299.

[109] M. Quigley, B. Gerkey, K. Conley, J. Faust, T. Foote, J. Leibs, E. Berger, R. Wheeler, and A. Ng, "Ros: an open-source robot operating system," in *Proc. of the IEEE Intl. Conf. on Robotics and Automation (ICRA) Workshop on Open Source Robotics*, Kobe, Japan, May 2009.

[110] A. Newell, K. Yang, and J. Deng, "Stacked hourglass networks for human pose estimation," in *Computer Vision – ECCV 2016*, B. Leibe, J. Matas, N. Sebe, and M. Welling, Eds. Cham: Springer International Publishing, 2016, pp. 483–499.

[111] J. Tompson, R. Goroshin, A. Jain, Y. LeCun, and C. Bregler, "Efficient object localization using convolutional networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 648–656.

[112] A. Nibali, Z. He, S. Morgan, and L. Prendergast, "3d human pose estimation with 2d marginal heatmaps," in *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 2019, pp. 1477–1485.

[113] J. Martinez, R. Hossain, J. Romero, and J. J. Little, "A simple yet effective baseline for 3d human pose estimation," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 2640–2649.

[114] S.-E. Wei, V. Ramakrishna, T. Kanade, and Y. Sheikh, "Convolutional pose machines," in *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 2016, pp. 4724–4732.

[115] S. Moccia, L. Migliorelli, V. Carnielli, and E. Frontoni, "Preterm infants' pose estimation with spatio-temporal features," *IEEE Transactions on Biomedical Engineering*, vol. 67, no. 8, pp. 2370–2380, 2020.

[116] X. Sun, B. Xiao, F. Wei, S. Liang, and Y. Wei, "Integral human pose regression," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 529–545.

[117] B. Artacho and A. Savakis, "Unipose: Unified human pose estimation in single images and videos," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 7035–7044.

[118] M. Tan and Q. V. Le, "Efficientnet: Rethinking model scaling for convolutional neural networks," in *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, ser. Proceedings of Machine Learning Research, K. Chaudhuri and R. Salakhutdinov, Eds., vol. 97. PMLR, 2019, pp. 6105–6114.

[119] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 801–818.

[120] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2980–2988.

[121] A. Martínez-González, M. Villamizar, O. Canévet, and J.-M. Odobez, "Residual pose: A decoupled approach for depth-based 3d human pose estimation," in *IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2020.

[122] C. Beaman, C. Peterson, R. Neptune, and S. Kautz, "Differences in self-selected and fastest-comfortable walking in post-stroke hemiparetic persons," *Gait & Posture*, vol. 31, no. 3, pp. 311 – 316, 2010.

[123] C. Shorten and T. Khoshgoftaar, "A survey on image data augmentation for deep learning," *Journal of Big Data*, vol. 6, pp. 1–48, 2019.

[124] I. Sárándi, T. Linder, K. O. Arras, and B. Leibe, "How robust is 3d human pose estimation to occlusion?" in *IROS Workshop - Robotic Co-workers 4.0*, 2018.

[125] L. Zhao, X. Peng, Y. Tian, M. Kapadia, and D. N. Metaxas, "Semantic graph convolutional networks for 3d human pose regression," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 3425–3435.

[126] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L. Chen, "Mobilenetv2: Inverted residuals and linear bottlenecks," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018, pp. 4510–4520.

[127] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 7132–7141.

[128] H. Joo, T. Simon, X. Li, H. Liu, L. Tan, L. Gui, S. Banerjee, T. S. Godisart, B. Nabbe, I. Matthews, T. Kanade, S. Nobuhara, and Y. Sheikh, "Panoptic studio: A massively multiview system for social interaction capture," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017.

[129] P. Enderby, "Disorders of communication: dysarthria," *Handbook of Clinical Neurology*, vol. 110, pp. 273–281, 2013.

[130] J. Lee, A. Madhavan, E. Krajewski, and S. Lingenfelter, "Assessment of dysarthria and dysphagia in patients with amyotrophic lateral sclerosis: Review of the current evidence," *Muscle & Nerve*, 2021.

[131] P. Rong, Y. Yunusova, J. Wang, and J. R. Green, "Predicting early bulbar decline in amyotrophic lateral sclerosis: A speech subsystem approach," *Behavioural Neurology*, vol. 2015, 2015.

[132] G. Defazio, M. Guerrieri, D. Liuzzi, A. F. Gigante, and V. Di Nicola, "Assessment of voice and speech symptoms in early parkinson's disease by the robertson dysarthria profile," *Neurological Sciences*, vol. 37, no. 3, pp. 443–449, 2016.

[133] F. Fussi, A. Cantagallo, and L. Bertozzini, *Profilo di valutazione della disartria: adattamento italiano del test di Robertson, raccolta di dati normativi e linee di trattamento*. Omega, 1999.

[134] B. J. Perry, R. Martino, Y. Yunusova, E. K. Plowman, and J. R. Green, "Lingual and jaw kinematic abnormalities precede speech and swallowing impairments in als," *Dysphagia*, vol. 33, no. 6, pp. 840–847, 2018.

[135] A. Bandini, S. Rezaei, D. L. Guarín, M. Kulkarni, D. Lim, M. I. Boulos, L. Zinman, Y. Yunusova, and B. Taati, "A new dataset for facial motion analysis in individuals with neurological disorders," *IEEE Journal of Biomedical and Health Informatics*, vol. 25, no. 4, pp. 1111–1119, 2020.

[136] M. Macedo, A. Candeias, and M. Marques, "Motion analysis for people with cerebral palsy: A vision based approach," in *2019 IEEE 16th International Conference on Rehabilitation Robotics*. IEEE, 2019, pp. 40–45.

[137] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask r-cnn," in *IEEE International Conference on Computer Vision*, 2017, pp. 2961–2969.

[138] C. Sagonas, E. Antonakos, G. Tzimiropoulos, S. Zafeiriou, and M. Pantic, "300 faces in-the-wild challenge: Database and results," *Image and Vision Computing*, vol. 47, pp. 3–18, 2016.