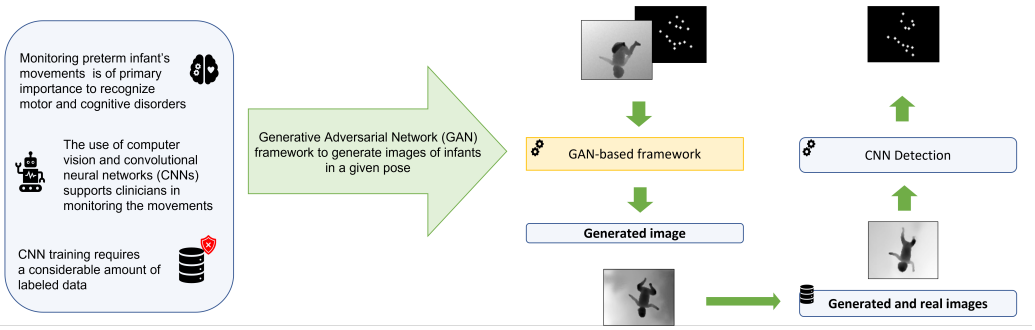
Generating depth images of preterm infants in given poses using GANs

Graphical Abstract

Generating depth images of preterm infants in given poses using GANs

Giuseppe Pio Cannata, Lucia Migliorelli, Adriano Mancini, Emanuele Frontoni, Rocco Pietrini, Sara Moccia



Highlights

Generating depth images of preterm infants in given poses using GANs

Giuseppe Pio Cannata, Lucia Migliorelli, Adriano Mancini, Emanuele Frontoni, Rocco Pietrini, Sara Moccia

- Datasets lack limits the spread of deep-learning algorithms to monitor preterm infants
- We propose a GAN framework to generate realistic depth images of preterm infants
- Validation is performed on the Moving INfants In RGB-D dataset with 12 infants
- Quantitative and qualitative evaluation of the framework shows promising results

Generating depth images of preterm infants in given poses using GANs

Giuseppe Pio Cannata[a,**], Lucia Migliorelli[a,**], Adriano Mancini[a], Emanuele Frontoni[b], Rocco Pietrini[a], Sara Moccia[c]

[a] *Department of Information Engineering, Universita' Politecnica delle Marche,*
[b] *Department of Political Science, Communication and International Relations, Universita' degli Studi di Macerata,*
[c] *The BioRobotics Institute and Department of Excellence in Robotics and AI, Scuola Superiore Sant'Anna,*

Abstract

Background and objectives: The use of deep learning for preterm infant's movement monitoring has the potential to support clinicians in early recognizing motor and behavioural disorders. The development of deep learning algorithms is, however, hampered by the lack of publicly available annotated datasets. ***Methods:*** To mitigate the issue, this paper presents a Generative Adversarial Network-based framework to generate images of preterm infants in a given pose. The framework consists of a bibranch encoder and a conditional Generative Adversarial Network, to generate a rough image and a refined version of it, respectively. ***Results:*** Evaluation was performed on the Moving INfants In RGB-D dataset which has 12000 depth frames from 12 preterm infants. A low Fréchet inception distance (142.9) and an inception score (2.8) close to that of real-image distribution (2.6) are obtained. The results achieved show the potentiality of the framework in generating realistic depth images of preterm infants in a given pose. ***Conclusions:*** Pursuing research on the generation of new data may enable researchers to propose increasingly advanced and effective deep learning-based monitoring systems.

Keywords: Depth-Image Generation, Generative Adversarial Networks, Preterm Infants' Monitoring

*Corresponding author: l.migliorelli@staff.univpm.it
**The authors equally contributed to the paper draft

1. Introduction

Preterm birth is defined by the World Health Organisation (WHO) as a birth occurring before the end of the 37th gestational week. This event affects 15 million newborns each year and its incidence tends to increase. Preterm birth occurs for a variety of reasons (the pregnant woman's Sars-Cov-2 infection is one of the major implications of preterm birth to date) and its consequences are the leading cause of death for children under 5 years of age [1].

Preterm birth, particularly extremely one, entails organs' immaturity, which causes infants' difficulties in coping with the extra-uterine environment and possible impairment of neurodevelopmental functioning. Timely identifying preterm infants at risk of developing severe neurological disorders is still an open challenge in clinics. The assessment of preterm infants' spontaneous movements, i.e., general movements, has a crucial diagnostic and prognostic role [2]. However, despite its clinical relevance, movement evaluation is sporadic and qualitative, as it mostly relies on rating scales following observation of infants' by clinicians in neonatal intensive care units (NICUs) [3].

To support NICU clinicians, over the years several computer-aided movement monitoring systems have been proposed in the literature [4]. Recently, vision systems have drawn the attention of the researchers [5], [6], [7], [8]. Indeed, these contactless systems do not influence infant's free movement neither cause discomfort. At the same time, they leave clinicians and parents free to interact with the infants [4].

These movement monitoring systems rely on deep learning (DL) to predict infants' pose or silhouette, tackling challenges such as high intra- and inter-infant variability in terms of body size and movement patterns [9].

The deployment of these DL methodologies for multimedia data analysis in the actual clinical practice is, however, hampered by the lack of large publicly available datasets for algorithm training. This data shortage may be explained considering that image collection in NICUs is hard -particularly during the pandemic period- and labelling is a tedious procedure that requires the supervision of expert clinicians require the supervision of experienced clinicians, whose efforts are currently deployed in the fight against the pandemic [10]. Furthermore, collecting videos of infants exhibiting patho-

logical movement patterns is not trivial: clinicians should review hours of recordings, select frames with general movements of interest and categorise them [11].

Over the years, we worked on implementing DL algorithms to preterm infants' limb-movement monitoring [12], [6], [5], [7], and often faced issues relevant to datasets that are not fully representative of the variability of preterm infants' movement. To attenuate this issue, in this work we present a generative adversarial network (GAN)-based framework to generate depth images depicting preterm infants in given poses. With our framework, issues relevant to image acquisition and labeling may be, at least partially, mitigated. The contribution of this work may be summarized as follows:

1. *Generation of depth images of preterm infants in given poses for supporting research in the field of DL monitoring systems for movement assessment.* Development of a novel GAN-based framework trained to generate depth images of infants in a given pose. Depth images are analyzed to attenuate privacy concerns (Sec. 2).
2. *Validation on a publicly available dataset.* A comprehensive study is conducted using 12000 depth frames acquired in the actual clinical practice to experimentally investigate our method. The dataset used is public to ensure fair comparisons (Sec. 3).

To foster research in the field, the code of the proposed GAN-based framework is available online[1].

1.1. State of the art

Extensive literature on GANs today exists [13] but few efforts were spent in the field of preterm infants. In [14], GANs are used to generate labeled RGB images for pedestrian-detection task. This is a two-steps methodology: (i) given an RGB image, pedestrians are localized and their bounding boxes are replaced by random noise, (ii) the new image is fed to a generator trained to replace the noise with a new pedestrian.

The work in [15] proposes a pedestrians' movements generation approach using GANs. A first network generates poses from input noise in the form of pedestrians' skeletons. A second generator is fed with these poses to output a realistic sequence of movements. A discriminator recognises whether the

[1] `https://vrai-group.github.io/guided-infant-generation`

produced sequence is realistic or not. The approach is tested on publicly available datasets (i.e., the Caltech Pedestrian dataset [16], Joint Attention in Autonomus Driving dataset [17] the Daimler dataset [18]).

A relevant approach is proposed in [19]. The authors explore the use of generative models to generate RGB images of pedestrians in given poses. An autoencoder generates rough images that are then refined and enriched in details via a conditional Deep Convolutional GAN (cDCGAN). The approach is validated on two publicly available datasets: the Market-1501 [20] and DeepFashion [21].

The only approach in the literature to infants' images generation is presented in [22]. The approach is mainly conceived for adult-image generation using the Human 3.6 dataset [23] and only preliminary results for infants are shown. The authors proposes an autoencoder architecture that, from an RGB image depicting a person and a target-pose image, generates a new RGB image of the same person in the target pose. Here, a main limitation lies in the low quality of the generated infants' images, which are poor in those details that are needed for monitoring application in clinical scenarios. Examples include limb extremities (hands and feet), which are blurred.

To tackle the limitation in [22] and inspired by [19], in this work we propose a GAN-based framework for generating depth images of infants in desired poses. Our approach shares a framework similar to the one proposed in [19] to ensure images rich in fine details. It should be noted that we chose to generate depth images because they preserve privacy compared to RGB ones, as each pixel in the depth image encodes the distance from the camera.

2. Methods

Figure 1 shows our GAN-based framework, which consists of a double-branch convolutional autoencoder (G_1) and a cDCGAN. As shown in [19], using such a framework, instead of a single autoencoder as in [22], allows to generate more realistic images.

Our GAN-based framework is fed with: (i) a depth input image of a preterm infant (I_C), which acts as condition image, and (ii) a target pose (P_T). P_T is a stack of N images, where each image is a keypoint mask built, as explained in Sec. 2.1, from a depth image (I_T) of a different infant. The purpose of our framework is to translate the infant depicted in I_C into the target pose P_T.

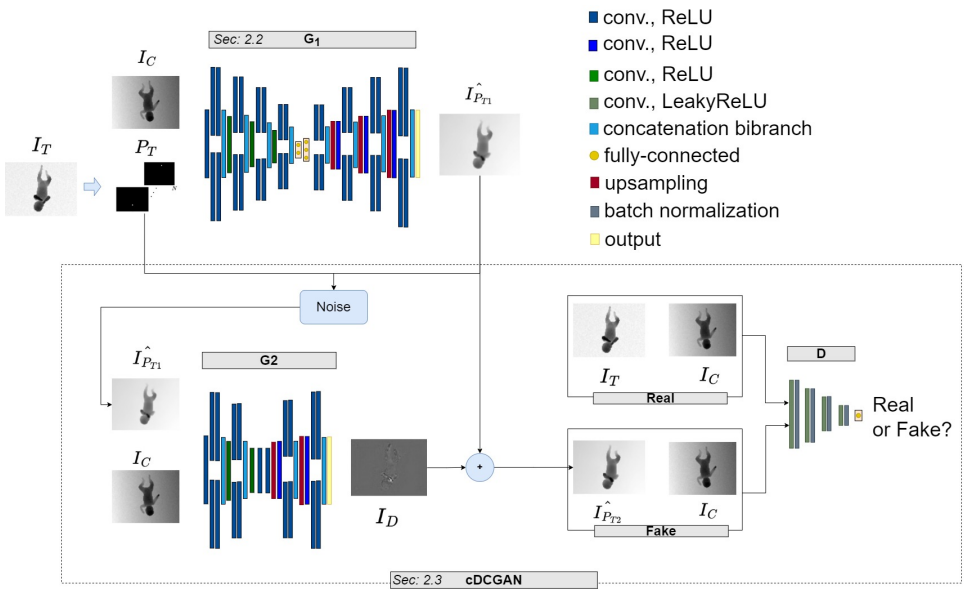


Figure 1: Workflow of the proposed generative adversarial network (GAN)-based framework to generate depth images of preterm infants with a given poses. Acronyms are reported in Table 1.

Table 1: Acronyms used in Sec. 2.

cDCGAN	Conditional deep convolutional GAN
D	Discriminator
G_1	Double branch convolutional autoencoder
G_2	Double branch convolutional autoencoder in cDCGAN
GAN	Generative adversarial network
H	Height of I_C
I_C	Input conditional depth image
I_D	Difference map in output from G_2
$\hat{I_{P_{T1}}}$	Output of G_1
$\hat{I_{P_{T2}}}$	Output of the proposed GAN-based framework
I_T	Depth image from a different infant
P_T	Target pose
N	Number of keypoints
W	Width of I_C

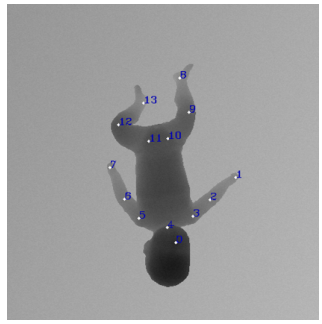


Figure 2: Top-view depth image with 14 infants' keypoints annotation is shown.

The choice of having two different infants for I_C and I_T is driven by the need of maximising data variability both in terms of movement patterns and infants' sizes.

In our framework, G_1 generates a rough image ($\hat{I_{P_{T1}}}$) of the infant in I_C with the desired pose P_T.

The output of G_1 is fed to the cDCGAN, which consists of a double-branch autoencoder generator (G_2) and a discriminator (D), where D is used during the training phase only.

G_2 and D are trained jointly to obtain a refined version ($\hat{I_{P_{T2}}}$) of $\hat{I_{P_{T1}}}$. In detail, G_2 outputs a difference map (I_D) with infant's fine details. The I_D is added to $\hat{I_{P_{T1}}}$ as to obtain $\hat{I_{P_{T2}}}$.

D is trained to classify the pairs (I_T, I_C) as real and ($\hat{I_{P_{T2}}}$, I_C) as fake, respectively. For D to recognise ($\hat{I_{P_{T2}}}$, I_C) as real, G_2 tries to produce an I_D as rich as possible in terms of infant's details.

Opposite to [19], we chose a double-branch architecture for G_1 and G_2 inspired by our previous work on infants' pose estimation [5], where the double-branch architecture allowed parallel processing of joints and joint connections, improving pose-estimation performance.

2.1. Data Preparation

Following our previous work [5], we modeled infant's pose as a set of 12 connected keypoints for limbs' joints, and 2 keypoints for neck and head, as shown in Fig. 2, for a total of $N = 14$ keypoints. Starting from this keypoint model, P_T was made of 14 binary masks with size HxW. Each mask refers to a single keypoint and was obtained masking all the pixels lying in a circle of a defined radius (r_k) centered at the keypoint.

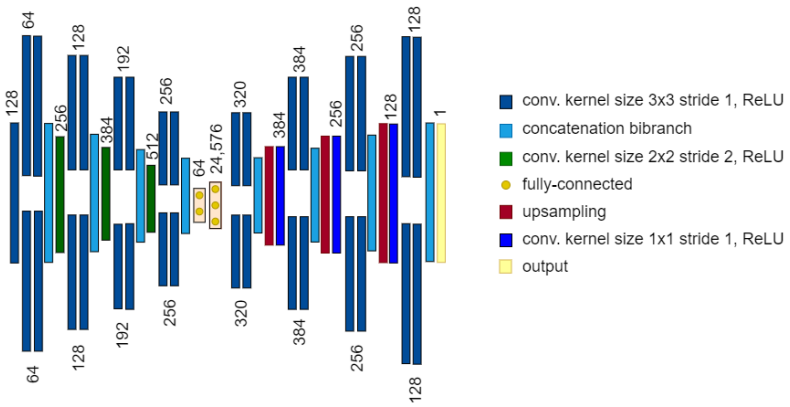


Figure 3: The G_1 architecture.

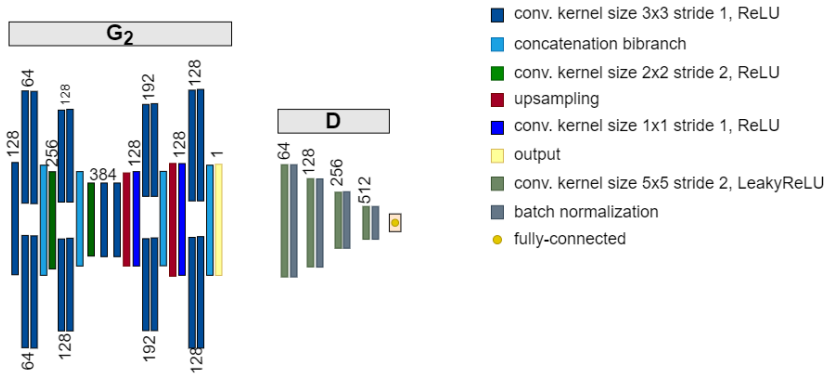


Figure 4: The cDCGAN architecture.

Table 2: G_1 architecture specification.

Name	Kernel (Size / Stride)	Channels
Downsampling path		
Input	–	15
Block 1 - Conv. layer	3x3 / 1	128
Block 1 - Branch 1	3x3 / 1	64
	3x3 / 1	64
Block 1 - Branch 2	3x3 / 1	64
	3x3 / 1	64
Block 1 - Concatenation	–	128
Block 2 - Conv. layer	2x2 / 2	256
Block 2 - Branch 1	3x3 / 1	128
	3x3 / 1	128
Block 2 - Branch 2	3x3 / 1	128
	3x3 / 1	128
Block 2 - Concatenation	–	256
Block 3 - Conv. layer	2x2 / 2	384
Block 3 - Branch 1	3x3 / 1	192
	3x3 / 1	192
Block 3 - Branch 2	3x3 / 1	192
	3x3 / 1	192
Block 2 - Concatenation	–	384
Block 4 - Conv. layer	2x2 / 2	512
Block 4 - Branch 1	3x3 / 1	256
	3x3 / 1	256
Block 4 - Branch 2	3x3 / 1	256
	3x3 / 1	256
Block 2 - Concatenation	–	512
Bridge		
Fully - connected 1	–	64
Fully - connected 2	–	12x16x128
Upsampling path		
Block 5 - Branch 1	3x3 / 1	320
	3x3 / 1	320
Block 5 - Branch 2	3x3 / 1	320
	3x3 / 1	320
Block 5 - Concatenation	–	640
Block 5 - Upsampling layer	–	640
Block 6 - Conv. layer	1x1/1	384
Block 6 - Branch 1	3x3 / 1	384
	3x3 / 1	384
Block 6 - Branch 2	3x3 / 1	384
	3x3 / 1	384
Block 6 - Concatenation	–	768
Block 6 - Upsampling layer	–	768
Block 7 - Conv. layer	1x1/1	256
Block 7 - Branch 1	3x3 / 1	256
	3x3 / 1	256
Block 7 - Branch 2	3x3 / 1	256
	3x3 / 1	256
Block 7 - Concatenation	–	512
Block 7 - Upsampling layer	–	512
Block 8 - Conv. layer	1x1 / 1	128
Block 8 - Branch 1	3x3 / 1	128
	3x3 / 1	128
Block 8 - Branch 2	3x3 / 1	128
	3x3 / 1	128
Block 8 - Concatenation	–	256
Output	1x1/1	1

2.2. G_1: the first double-branch autoencoder

The architecture of G_1 is shown in Fig. 3 and described in details in Table 2.

The down-sampling path (*encoder*) of G_1 consists of 4 consecutive double-branch blocks. Apart from the first block where the first convolutional layer has kernel size 3x3 and stride 1, the subsequent blocks are built as follows:

1. One convolutional layer with F_i kernels, each with size 2x2, with a stride of 2.
2. Two branches with two convolutional layers with $F_i/2$ filters, kernel size 3x3 and stride 1.
3. A layer to concatenate the features maps in output from each branch.

The output of each convolutional layers is activated with the Rectified Linear Unit (ReLU) activation function. The F_i for each block is reported in details in Table 2.

As in [19], the encoder ends with two fully-connected layers aimed at empowering information exchange between distant body parts.

The number of blocks in the up-sampling path (decoder) is equal to that of the encoder, but each double-branch, except for the first block, has twice as many kernels as the corresponding block in the encoder. This provides the network with boosted generalisation This provides the network with boosted generalisation power [19].

2.3. cDCGAN: Conditional Deep Convolutional Generative Adversarial Network

G_2 shares the same architecture of G_1 except for the 2 fully-connected layers and the last two double-branch blocks of G_1. This architectural simplification is driven by the fact that G_2 needs to produce an image with only high-frequency details [19].

The architecture of D in Table 4 is made of 4 convolutional blocks followed by a fully-connected layer. The blocks are made of 2 convolutional layers with an increasing number of filters, from 64 to 512. The last layer is a binary classification layer with a neuron. Following standard guidelines for GANs [24], each convolutional kernel has size 5x5 and strides of 2, to reduce the feature-map size. Batch normalization is applied at the end of each convolutional block and LeakyReLU activation function with a value of

Table 3: G_2 architecture specification.

Name	Kernel (Size / Stride)	Channels
Downsampling path		
Input	–	15
Block 1 - Conv. layer	3x3 / 1	128
Block 1 - Branch 1	3x3 / 1	64
	3x3 / 1	64
Block 1 - Branch 2	3x3 / 1	64
	3x3 / 1	64
Block 1 - Concatenation	–	128
Block 2 - Conv. layer	2x2 / 2	256
Block 2 - Branch 1	3x3 / 1	128
	3x3 / 1	128
Block 2 - Branch 2	3x3 / 1	128
	3x3 / 1	128
Block 2 - Concatenation	–	256
Conv. layer	2x2 / 2	384
Conv. layer	3x3 / 1	384
Conv. layer	3x3 / 1	384
Upsampling layer	–	384
Upsampling path		
Block 3 - Conv. layer	–	128
Block 3 - Branch 1	3x3 / 1	192
	3x3 / 1	192
Block 3 - Branch 2	3x3 / 1	192
	3x3 / 1	192
Block 3 - Concatenation	–	384
Block 3 - Upsampling layer	–	384
Block 4 - Conv. layer	–	128
Block 4 - Branch 1	3x3 / 1	128
	3x3 / 1	128
Block 4 - Branch 2	3x3 / 1	128
	3x3 / 1	128
Block 4 - Concatenation	–	256
Output	1x1/1	1

Table 4: D architecture specification.

Name	Kernel (Size / Stride)	Channels
Input	–	1
Block 1 - Conv. layer	5x5 / 2	64
Block 1 - Batch Norm. layer		
Block 2 - Conv. layer	5x5 / 2	128
Block 2 - Batch Norm. layer		
Block 3 - Conv. layer	5x5 / 2	256
Block 3 - Batch Norm. layer		
Block 4 - Conv. layer	5x5 / 2	512
Block 4 - Batch Norm. layer		
Output		1

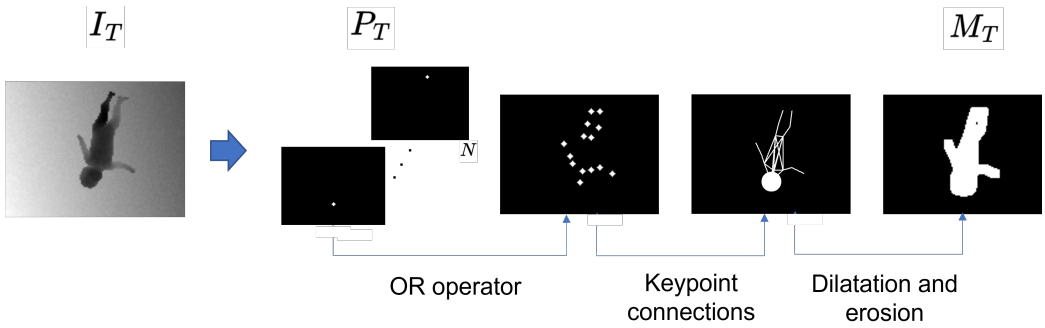


Figure 5: To compute the PoseMask Loss ($L^{G_1}_{PoseMask}$), infant's rough segmentation mask (M_T) is obtained from I_T using morphological operators.

negative slope coefficient equal to 0.2 is used. The neuron in the last layer is activated with a linear function.

While conducting our experiments we realized that G_2 produced a completely black I_D. This happened because the discriminator immediately classified ($\hat{I_{P_{T2}}}$, I_C) as a real pair, inducing, as a consequence, G_2 to not produce any refinement. However, the quality of $\hat{I_{P_{T1}}}$ was still too poor.

Thus, to facilitate D in recognizing ($\hat{I_{P_{T2}}}$, I_C) as fake, we added noise to $\hat{I_{P_{T1}}}$. This was done on-the-fly during the cDCGAN training adding random noise to each P_T for each batch.

2.4. Training protocol

To train G_1, we implemented a custom loss, i.e. the PoseMask Loss ($L^{G_1}_{PoseMask}$):

$$L^{G_1}_{PoseMask} = ||(\hat{I_{P_{T1}}} - I_T) \odot (1 + M_T)||_1 \tag{1}$$

where $\odot$ and M_T refers to pixel-wise multiplication and infant-body rough segmentation, respectively.

The $L^{G_1}_{PoseMask}$ is an $L1$ loss aimed at capturing global information of the target image I_T focusing on the infant's body, identified by M_T. With a view to obtain M_T, as shown in Fig. 5, all the P_T masks were joined together in a single image using the boolean OR operation. The radius (r_h) for the head keypoint mask was dilated to overlay the whole head. Each keypoint was

linked to its neighbours and morphological operations (dilation and erosion) were applied.

D was trained to output 1 and 0 for the real (I_C, I_T) and fake (I_C, $\hat{I_{P_{T2}}}$) pair, respectively, minimizing the adversarial loss (L^D_{adv}):

$$L^D_{adv} = L_{bce}(D(I_T, I_C), 1) + L_{bce}(D(\hat{I_{P_{T2}}}, I_C), 0) \tag{2}$$

where L_{bce} is the binary cross-entropy loss.

G_2 was trained to cheat D by improving the $\hat{I_{P_{T2}}}$ output minimizing the loss:

$$L^{G_2}_{adv} = L_{bce}(D(\hat{I_{P_{T2}}}, I_C), 1) + \lambda||(\hat{I_{P_{T2}}} - I_T) \odot (1 + M_T)||_1 \tag{3}$$

where λ is a constant parameter that was set experimentally. For training the proposed GAN-based framework, Adam was used as optimizer.

3. Experimental Protocol

3.1. Dataset

The dataset used in this study was the Moving INfants In RGB-D (MINI-RGBD) [25], a publicly available dataset of videos from 12 infants recorded in top-view mode. The dataset consists of 12000 depth frames (i.e., 1000 frames for infant) with resolution 480x640 pixels. To improve data variability by reducing the amount of similar movements, downsampling was performed every 5 frames, resulting in a total of 200 frames per infant. Such a downsampling procedure is in line with the infants' movement rate [26]. All frames were: (i) resized, with a nearest-neighbor interpolation technique, from 480x640 to 96x128 pixels to reduce the training time and amount of memory required and (ii) normalized to the intensity range [-1,1]. P_T was obtained considering r_k equal to 2 pixels.

The dataset was split using 10 infants to train and validate the framework and the remaining 2 for testing purposes. Each pair (I_C, I_T) was obtained without mixing infants from training, validation and test set to avoid possible biases [19]. The total number of pairs was 11.600 for training and validation and 400 for testing.

3.2. Training settings

For M_T, we set r_h equal to 40 pixels to fully overlay the head surface. We set the learning rate equal to 2e-5 for G_1, G_2 and D. The decay of the learning rate was used only for G_1, setting a step decay equal to 0.5 at each epoch.

We chose a batch size of 16, as a trade-off between training speed and memory constraints, and set the number of training epochs equal to 10 and 200 for G_1 and cDCGAN, respectively. The higher number of epochs for the cDCGAN was required due to the more complex architecture. After extensive tuning, we set λ equal to 10. To attenuate vanishing-gradient issues, the training of G_2 occurred during all multiple iterations of 3, while the training of D occurred during the remaining iterations [19].

In order to increase the variability of the training set, particularly in terms of infants' positions with respect to the camera field of view, on-the-fly data augmentation was performed. (i) Affine (rotation in range [-90,90] degrees, vertical shift in [-30,30] pixels and horizontal shift in [-10,10] pixels) and (ii) structural transformations, were applied during G_1 training.

3.3. Performance assessment and ablation study

Assessing the performance of GANs is still an open challenge in the literature [27]. In this work, we decided to compute the Inception Score (IS) [28] and Fréchet Inception Distance (FID) [29]. Lower values of FID mean better results. These metrics capture the quality (FID) and diversity (IS) of the generated images, and are widely used in the literature for evaluation of generated images [19, 30]. For the IS score on the generated distribution, we considered a value close to the IS score on the real distribution to be good.

These aforementioned metrics rely upon the Inception network pretrained on ImageNet. Considering that ImageNet dataset has RGB images, we replicated for 3 times the depth frame, obtaining a 3-channel image. We further resized it to 299x299 pixels (in line with pretraining), removed image mean and normalized by the standard deviation.

To focus the evaluation only on infants without considering the background, we further designed the Mask-FID and the Mask-IS score, which are computed from the generated image multiplied by the corresponding M_T.

Quantitative performance with the t-distributed stochastic neighbor embedding (t-sne) plot was computed too. This plot maps high-dimensional data by assigning each point a position in a two-dimensional map. To get the t-sne plot, each real and generated image was embedded with a feature

Table 5: Performed experiments, where *mono* refers to the implementation of monobranch blocks.

	G_1^{mono}	G_1	G_2^{mono}	G_2
E1 - Monobranch autoencoder	✓			
E2 - Bibranch autoencoder		✓		
E3 - Monobranch GAN-based framework [19]	✓		✓	
Proposed GAN-based framework		✓		✓

vector of size 512, obtained from the Global Average Pooling layer of VGG-16 pretrained on ImageNet. Principal Component Analysis (PCA) was applied on the 512 features retaining the first 50 principal components. For the t-sne plot we considered a perplexity and number of iteration of 50 and 6000, respectively, as done in [31].

We conducted ablation studies as shown in Table 5. The first experiment tested the performance of an autoencoder, as the one proposed in [22]. The architecture was the same presented in Sec. 2.2 but with monobranch blocks (i.e., each block of convolutions in consists of a single branch). We further considered the bibranch version of the autoencoder in E2 (i.e., $G1$). In E3, we implemented the monobranch version of our GAN-based framework which is the one proposed in [19].

4. Results

Table 6 shows the quantitative results in terms of IS and FID on test set for E1, E2, E3 and the proposed GAN-based framework. E1 achieved the IS value (2.79) closest to that of the real dataset, followed by our framework (2.85), E3 (2.88), and E2 (2.99). When considering Mask-IS, E3 achieved the best score (2.67), followed by E1 (2.65), the proposed framework (2.54) and E2 (2.47). The E1 (165.77) and E2 (179.20) achieved the highest values for FID, highlighting the benefit of the cDCGAN framework. The same trend was seen for Mask-FID. Particularly, the proposed GAN-based framework achieved a MASK-FID similar to that of [19], with a difference of 4.79.

Figure 6 shows the qualitative results for all the conducted experiments. When observing the figure our GAN-proposed framework generates images closest to the real ones and is able to reproduce the infant's upper and lower limbs with finer details with respect to the architectures of the other experiments.

Table 6: Quantitative metrics obtained with E2 and by the proposed GAN-based framework, where real and generated refer to the metrics computed on the real and generated images of the test set, respectively.

	IS (real)	IS (generated)	Mask-IS (real)	Mask-IS (generated)	FID	Mask-FID
E1	2.67	**2.79**	3.26	2.65	165.77	152.02
E2	2.67	2.99	3.26	2.47	179.20	148.23
E3	**2.67**	2.88	3.26	2.67	**126.45**	**125.39**
Proposed	2.67	2.85	3.26	2.54	142.94	130.18

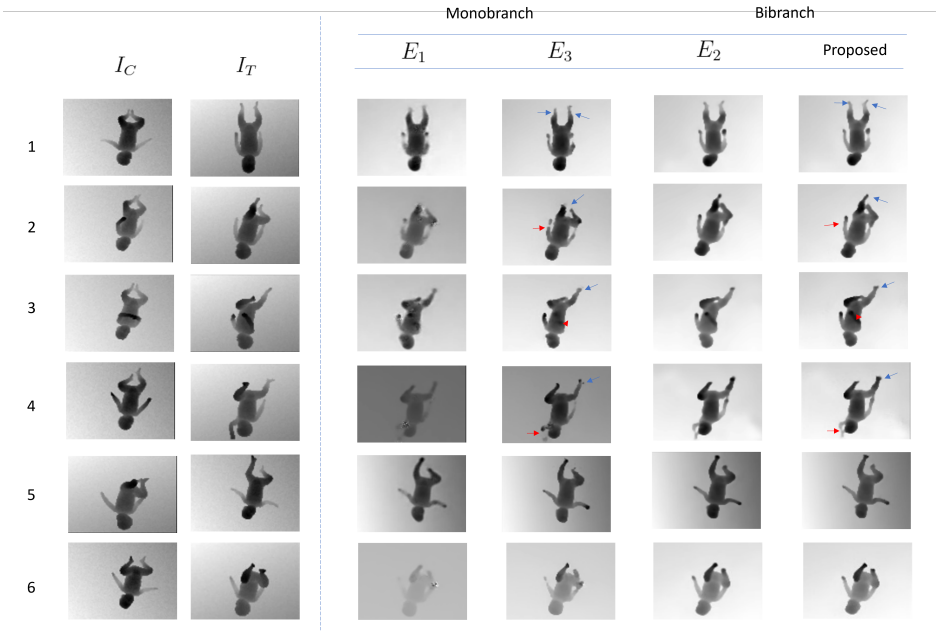


Figure 6: Qualitative results for input I_C. Results are shown for E1, E3, E2, and for the proposed framework in columns 3,4,5,6, respectively. In rows 1, 2, 3, arrows highlight main difference in the generation of lower- (blue) and upper-limb (red) details. Samples on rows 5 and 6 show that the depth information is more realistic for the bibranch implementation. I_T in column 2 is shown too as to clarify in which desired pose P_T the infant depicted in I_C should be translated.

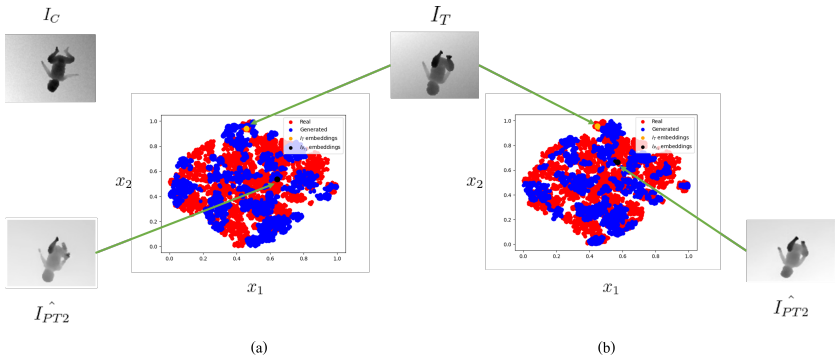


Figure 7: Comparison of real (red) and generated (blue) embeddings for the (a) monobranch (E3) and (b) the proposed GAN-based framework using the 2D t-sne plot. x_1 and x_2 are the two principal components. The embeddings of a sample I_T image and the generated $\hat{I_{P_{T2}}}$ in the same pose are highlighted, too.

In Fig. 7, the distribution for the generated and real embedded features is shown for both proposed GAN-based framework and E3. The embeddings of a sample I_T image are shown, too, along with the corresponding embeddings of $\hat{I_{P_{T2}}}$. For t-sne plot we selected a challenging body pose and assessed the Euclidean distance between the embedded features from our GAN-based framework and E3. The pose reflected a crouched infant and his/her legs are closer to the camera so the framework must be able to generate an image in which the lower limbs have a different pixels' values (e.g., appearing darker) with respect to the other body portions. The Euclidean distance in the feature space was equal to 0.28 for the proposed GAN framework while for the monobranch (i.e., E3) was equal to 0.44.

5. Discussion

Monitoring preterm infants' movement in NICU is crucial to early detect motor and behavioural disorders. For supporting clinicians a number of computer-assisted approaches for RGB-D image analysis based on supervised DL methodologies were proposed in literature [4]. All the approaches mainly rely upon annotated datasets which are limited in size and variability (e.g., infants' gestational weeks and clinical conditions).

Dealing with limited datasets may pose issues relevant to the reliability of such DL-based monitoring systems [32, 33, 34]. Biases such as (i) *label bias*, when a training set is not fully representative of the infants' movement-

pattern variability [35] and (ii) *cohort* and *minority bias*, when, due to the small dataset size, DL algorithms cannot generalise when deployed in the actual clinical practice [36], may hamper the robustness of DL-based monitoring system [37].

To mitigate the issue, in this work we proposed a framework for the generation of infants' depth images in desired poses. We worked with depth images over RGB ones to fully respect ward and infants' privacy.

The quantitative analysis (Table 6) we performed, despite being extensively used for evaluating GANs, may give misleading results for models trained on datasets other than ImageNet, and this is particularly true in our case where we work with one-channel images [27].

This aspect mainly involves the IS score which, unlike the FID and the t-sne, does not compare the generated distribution with the real one but relies only upon features extracted from a network pretrained on ImageNet. This, as stated in [19], may not highlight the actual quality of the generated images, as the network (i.e., Inception, in the case of IS) may be unsuitable to extract consistent features for our task. Therefore, we decided to support quantitative results with qualitative ones (Figure 6).

As showed in Table 6 (FID and Mask-FID columns) and Figure 6, including the cDCGAN (E3 and proposed framework) over using a single autoencoder improved the performance of image generation. This supports the considerations made in [19] in which the authors argued that the addition of the cDCGAN guarantees more realistic images.

Considering the t-sne plot in Fig. 7, both the monobranch (E3) and the proposed GAN-based framework generated embedded features lying in the same domain space of the real embedded features. When assessing the Euclidean distance in the features domain our GAN-based framework achieves the highest performance (i.e., the shortest Euclidean distance) meaning that its generated images are closer to real ones than that of the monobranch framework. This is further confirmed by observing Figure 6 rows 1, 2, 3, 4. As visible, the generated images with the proposed framework were richer in details than the ones from E3 and the other experiments, particularly for limbs. This ability to finely generate infants' limbs is relevant with the view to develop DL-based support systems for clinicians in NICUs. Indeed, the early diagnosis of neuromotor disorders depends on the assessment of infants' limbs movement [4].

A limitation of the proposed framework may be seen in the fact that we did not consider pathological poses since the dataset we used did not have

any. However, the purpose of this work was to investigate the feasibility of generating depth images of infants in given poses as a way to provide new data for training DL-based movement-monitoring systems for NICUs.

6. Conclusion

This paper proposed a GAN-based framework for generating depth images in a given pose, showing promising results on the publicly-available MINI-RGBD dataset [25]. Future research directions of the presented work include: the generation of temporal sequences over still frames and the subsequent clinical validation of the generated images and videos. We will develop a web application for enabling clinicians to draw custom poses (both pathological and non-pathological) as a prior to generate realistic depth images, similar to what is done with GauGAN[2]. These generated data will be used to test the performance of existing DL-based systems for preterm infants' movement monitoring [38].

Acknowledgments

Statement of Ethical Approval

This study did not need any ethical approval.

Conflict of interest statement

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] R. Wood, C. Sinnott, I. Goldfarb, M. Clapp, T. McElrath, S. Little, Preterm birth during the coronavirus disease 2019 (covid-19) pandemic in a large hospital system in the united states, Obstetrics and Gynecology 137 (3) (2021) 403.

[2] `http://gaugan.org/gaugan2/`

[2] J. Zhou, S. Li, L. Gu, X. Zhang, Z. Tang, General movement assessment is correlated with neonatal behavior neurological assessment/cerebral magnetic resonance imaging in preterm infants, Medicine 100 (37) (2021).

[3] A. Steiner, Bayley scales of infants development-II, Encyclopedia of Autism Spectrum Disorders (2021) 605–606.

[4] K. Raghuram, S. Orlandi, P. Church, T. Chau, E. Uleryk, P. Pechlivanoglou, 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 63 (6) (2021) 637–648.

[5] S. Moccia, L. Migliorelli, V. Carnielli, E. Frontoni, Preterm infants' pose estimation with spatio-temporal features, IEEE Transactions on Biomedical Engineering 67 (8) (2020) 2370–2380.

[6] L. Migliorelli, D. Berardini, F. Rossini, E. Frontoni, V. Carnielli, S. Moccia, Asymmetric three-dimensional convolutions for preterm infants' pose estimation, in: 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society, IEEE, 2021, pp. 3021–3024.

[7] L. Migliorelli, E. Frontoni, S. Appugliese, G. P. Cannata, V. Carnielli, 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 & Biology Society, IEEE, 2021, pp. 3013–3016.

[8] V. Marchi, A. Hakala, A. Knight, F. D'Acunto, M. L. Scattoni, A. Guzzetta, S. Vanhatalo, Automated pose estimation captures key aspects of general movements at eight to 17 weeks from conventional videos, Acta Paediatrica 108 (10) (2019) 1817–1824.

[9] I. Bernhardt, M. Marbacher, R. Hilfiker, 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 87 (9) (2011) 633–639.

[10] J. Frazer, P. Notin, M. Dias, A. Gomez, J. K. Min, K. Brock, Y. Gal, D. S. Marks, Disease variant prediction with deep generative models of evolutionary data, Nature 599 (7883) (2021) 91–95.

[11] M. Porro, C. Fontana, M. L. Giannì, N. Pesenti, T. Boggini, A. De Carli, G. De Bon, G. Lucco, F. Mosca, M. Fumagalli, et al., Early detection of general movements trajectories in very low birth weight infants, Scientific Reports 10 (1) (2020) 1–7.

[12] S. Moccia, L. Migliorelli, R. Pietrini, 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, 2019, pp. 1–7.

[13] V. Sampath, I. Maurtua, J. J. Aguilar Martín, A. Gutierrez, A survey on generative adversarial networks for imbalance problems in computer vision tasks, Journal of Big Data 8 (1) (2021) 1–59.

[14] X. Ouyang, Y. Cheng, Y. Jiang, C.-L. Li, P. Zhou, Pedestrian-synthesis-gan: Generating pedestrian data in real scene and beyond, arXiv preprint arXiv:1804.02047 (2018).

[15] J. Spooner, V. Palade, M. Cheah, S. Kanarachos, A. Daneshkhah, Generation of pedestrian crossing scenarios using ped-cross generative adversarial network, Applied Sciences 11 (2) (2021) 471.

[16] P. Dollár, C. Wojek, B. Schiele, P. Perona, Pedestrian detection: A benchmark, in: 2009 IEEE conference on computer vision and pattern recognition, IEEE, 2009, pp. 304–311.

[17] A. Rasouli, I. Kotseruba, J. K. Tsotsos, Are they going to cross? a benchmark dataset and baseline for pedestrian crosswalk behavior, in: Proceedings of the IEEE International Conference on Computer Vision Workshops, 2017, pp. 206–213.

[18] N. Schneider, D. M. Gavrila, Pedestrian path prediction with recursive bayesian filters: A comparative study, in: German Conference on Pattern Recognition, Springer, 2013, pp. 174–183.

[19] L. Ma, X. Jia, Q. Sun, B. Schiele, T. Tuytelaars, L. Van Gool, Pose guided person image generation, Advances in Neural Information Processing Systems 30 (2017).

[20] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, Q. Tian, Scalable person re-identification: A benchmark, in: Proceedings of the IEEE international conference on computer vision, 2015, pp. 1116–1124.

[21] Z. Liu, P. Luo, S. Qiu, X. Wang, X. Tang, Deepfashion: Powering robust clothes recognition and retrieval with rich annotations, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 1096–1104.

[22] L. Schmidtke, A. Vlontzos, S. Ellershaw, A. Lukens, T. Arichi, B. Kainz, Unsupervised human pose estimation through transforming shape templates, in: IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 2484–2494.

[23] C. Ionescu, D. Papava, V. Olaru, C. Sminchisescu, Human3.6m: Large scale datasets and predictive methods for 3d human sensing in natural environments, IEEE Transactions on Pattern Analysis and Machine Intelligence 36 (7) (2014) 1325–1339.

[24] S. C. Alec Radford, Luke Metz, Unsupervised representation learning with deep convolutional generative adversarial network networks, International Conference on Learning Representations (2016) 3.

[25] N. Hesse, C. Bodensteiner, M. Arens, U. G. Hofmann, R. Weinberger, A. S. Schroeder, Computer vision for medical infant motion analysis: State of the art and RGB-D data set, in: Computer Vision - ECCV 2018 Workshops, Springer International Publishing, 2018.

[26] B. Fallang, O. D. Saugstad, J. Grøgaard, M. Hadders-Algra, Kinematic quality of reaching movements in preterm infants, Pediatric Research 53 (5) (2003) 836. doi:10.1203/01.PDR.0000058925.94994.BC.

[27] S. Barratt, R. Sharma, A note on the inception score, arXiv preprint arXiv:1801.01973 (2018).

[28] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, X. Chen, Improved techniques for training gans, Advances in Neural Information Processing Systems 29 (2016) 2234–2242.

[29] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, S. Hochreiter, Gans trained by a two time-scale update rule converge to a local nash equilibrium, Advances in Neural Information Processing Systems 30 (2017).

[30] T. Karras, S. Laine, T. Aila, A style-based generator architecture for generative adversarial networks, in: IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 4401–4410.

[31] V. Costa, N. Lourenço, J. Correia, P. Machado, Demonstrating the evolution of gans through t-sne, in: International Conference on the Applications of Evolutionary Computation (Part of EvoStar), Springer, 2021, pp. 618–633.

[32] A. Marzullo, S. Moccia, M. Catellani, F. Calimeri, E. De Momi, Towards realistic laparoscopic image generation using image-domain translation, Computer Methods and Programs in Biomedicine 200 (2021) 105834.

[33] E. Ovalle-Magallanes, J. G. Avina-Cervantes, I. Cruz-Aceves, J. Ruiz-Pinales, Improving convolutional neural network learning based on a hierarchical bezier generative model for stenosis detection in x-ray images, Computer Methods and Programs in Biomedicine (2022) 106767.

[34] N. J. Cronin, T. Finni, O. Seynnes, Using deep learning to generate synthetic b-mode musculoskeletal ultrasound images, Computer methods and programs in biomedicine 196 (2020) 105583.

[35] H. Jiang, O. Nachum, Identifying and correcting label bias in machine learning, in: International Conference on Artificial Intelligence and Statistics, PMLR, 2020, pp. 702–712.

[36] B. Heaton, K. M. Applebaum, K. J. Rothman, D. R. Brooks, T. Heeren, T. Dietrich, R. I. Garcia, The influence of prevalent cohort bias in the association between periodontal disease progression and incident coronary heart disease, Annals of Epidemiology 24 (10) (2014) 741–746.

[37] A. Rajkomar, M. Hardt, M. D. Howell, G. Corrado, M. H. Chin, Ensuring fairness in machine learning to advance health equity, Annals of internal medicine 169 (12) (2018) 866–872.

[38] 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, Expert Systems with Applications (2022) 117458.