A Monitoring Support System Based on HRNet+DEKR for Neonatal Intensive Care Units

Greta Di Marino[1]

Lucia Migliorelli[2]

Sara Moccia[3]

Claudio Sirocchi[2]

Matteo Lorenzo Bramucci[2]

Jacopo Carpaneto[1]

Alessandro Cacciatore[4]

Daniele Berardini[2]

[1] The BioRobotics Institute
Scuola Superiore Sant'Anna
Pisa, IT

[2] Department of Information Engineering
Università Politecnica delle Marche
Ancona, IT

[3] Department of Innovative Technologies
in Medicine and Dentistry
Università degli Studi 'G. d'Annunzio'
Chieti - Pescara
Chieti, IT

[4] Department of Humanities, Languages,
Mediation, History, Arts, Philosophy
Università di Macerata
Macerata, IT

Abstract

Preterm birth can lead to neurological disorders such as behavioral abnormalities, cognitive impairments, and delayed language development, highlighting the need for early diagnosis and intervention. Currently, the detection of these disorders relies on the qualitative assessment of General Movements (GMs) by clinicians, which is subjective and prone to variability. To address this, we propose a Deep Learning (DL) pipeline that leverages the High-Resolution Network coupled with a Disentangled Keypoint Regression (HRNet+DEKR) model for 2D pose estimation, aiming to enhance the accuracy and objectivity of GM assessments. Our approach uses a disentangled keypoint regression mechanism to detect 14 keypoints, focusing on pose estimation of limb as a preliminary step toward the evaluation of the potential neurological disorder. Trained and validated on depth images from the expanded BabyPose dataset, our model achieved an Average Precision (AP) of 0.975, average recall (AR) of 0.985, and introduced the novel Limb Overlap Score (LOS) as an evaluation metric, achieving an LOS of 0.992. Additionally, the HRNet+DEKR model demonstrated encouraging performance in multi-person scenarios, with promising qualitative results. These advancements pave the way for real-time clinical use, enabling more efficient and objective monitoring of infant neurodevelopment.

1 Introduction

The World Health Organisation (WHO) recently released a report that states that every year, almost 1 in 10 infants are born prematurely—that is, before 37 weeks of gestation[1]. Preterm birth affects the neurodevelopment of an infant, which can cause issues such as behavioral abnormalities, cognitive impairments, and delayed language development [13]. Early interventions are crucial for reducing the risk of physical disabilities and supporting the overall well-being of infants [3].

In the current clinical practice, the assessment of General Movements (GMs), i.e., infants' spontaneous movements, is used as an indicator for the early detection of neurobehavioural disorders [1]. GMs assessment involves trained clinicians visually inspecting infants in the Neonatal Intensive Care Unit (NICU). This approach may be, however, qualitative, discontinuous, and susceptible to variation among clinicians [9].

Several computer-assisted systems have been suggested in the literature as a solution to mitigate these issues, as reported by a comprehensive review on the topic [11]. The review concludes that using RGB-D cameras to support the assessment of infants' movement is an effective solution, as the cameras do not interfere with the infant's spontaneous movements or the clinical activities of healthcare providers. Additionally, pose estimation from these videos is frequently used as a preliminary step in the movement assessment process.

Although human pose estimation has been widely studied [14], progress in infant pose estimation has been slower due to several unique challenges. These include the limited availability of datasets to train algorithms, privacy concerns involving patients, healthcare staff, and caregivers, as well as the short time windows available to record infant movements, since infants spend much of their time sleeping or receiving care and treatments.

In response to these challenges, different approaches have been investigated. In [5] RGB cameras and Deep Learning (DL) algorithms were employed to monitor GMs in the NICU, using a two-step framework with a custom-built Convolutional Neural Network (CNN)-based pose estimation model and a subsequent movement analysis model to classify normal from cramped-synchronised GMs. This offers a promising sensor-free approach for early cerebral palsy risk assessment, although further validation with larger datasets is required, as 620 frames were used for training and 140 for testing to assess pose estimation. In [10], a semi-supervised learning framework, SiamParseNet (SPN), is introduced. SPN integrates both body parsing and pose estimation using a siamese network architecture. It addresses issues such as occlusions and limited labelled data, thus enhancing the assessment of movement from RGB videos. While both approaches [5, 10] show potential for non-invasive infant assessments, they may also raise privacy concerns in clinical settings due to the sensitive information captured. In [4], the authors implemented a transformer-based approach to estimate infant's pose using the Simultaneously collected multimodal Mannequin Lying pose dataset. Their method yielded promising results by addressing the challenge of keypoint occlusions. However, the dataset consisted of video recordings of covered mannequins, which were used to simulate infants wrapped in blankets.

Recently, sustainable and efficient DL models have been developed to further improve infant pose estimation while addressing privacy concerns by using depth data. For example, in [8], the TwinEDA model leverages depth images from the BabyPose dataset [6].

Recent advancements in pose estimation for infants have demonstrated the effectiveness of High-Resolution Network (HRNet)-based models. Building on the framework introduced

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

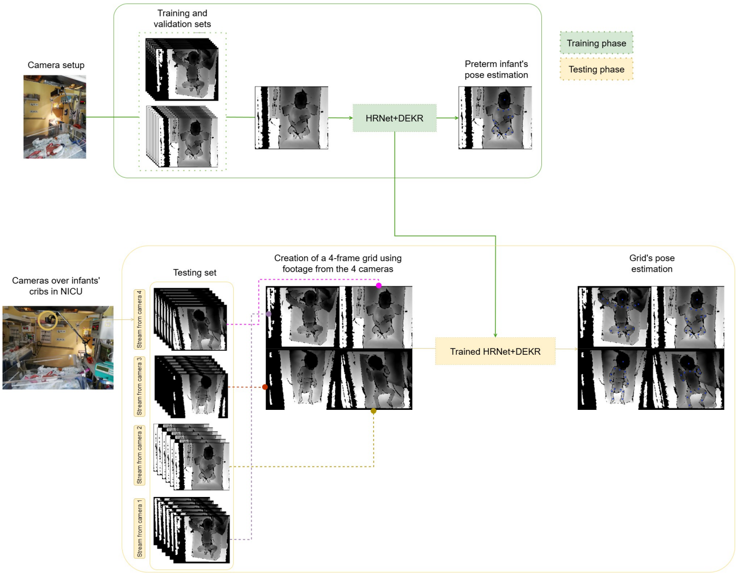

Figure 1: Pipeline for preterm infants' pose estimation from depth images. In the training phase, cameras placed over the infants' cribs (from [6, 7]) capture depth images, which are then split into training and validation sets. The High-Resolution Network + Disentangled Keypoint Regression (HRNet+DEKR) model is trained on these images to estimate the infant's pose. In the testing phase, footage from four different cameras is combined to create a 4-frame grid (2x2). The trained HRNet+DEKR model is applied to this grid to estimate the infants' pose.

in [12], which used HRNet, HigherHRNet, and DarkPose trained on over 88,000 images, our work proposes a multi-pose estimation approach using depth images from [6]. This system enables simultaneous monitoring of multiple infants in the NICU with a single model, enhancing computational efficiency. Following [12], for our purposes we leverage HRNet's integrated with a Disentangled Keypoint Regression (HRNet+DEKR) [2]. We show our pipeline in Figure 1.

2 Materials and methods

In our pipeline's training phase, we train and validate the HRNet+DEKR architecture to regress an infant pose from a single depth frame. As shown in Figure 2 (left), HRNet maintains high resolution throughout data flow while introducing parallel, lower-resolution levels that interact and exchange information. Unlike most CNNs, HRNet processes lower-resolution feature maps at the same depth. Each stage of HRNet has a high-resolution subnetworks and increasingly more subnetworks that handle incrementally lower resolutions. Information exchange across subnetworks is handled by exchange units using upsampling or downsampling to maintain feature coherency across resolutions. Downsampling is achieved via strided 3×3 convolutions, whereas upsampling employs nearest-neighbor sampling followed by a 1×1 convolution to align channel counts.

A key feature of the HRNet+DEKR architecture is its disentangled keypoint regression

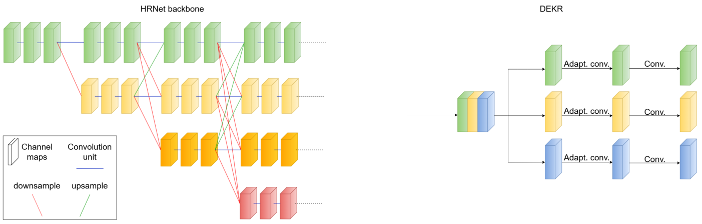

Figure 2: (Left) High-Resolution Network (HRNet) backbone structure: multi-branch and fusion features. The feature maps are processed at different resolutions—from green (high resolution) to red (low resolution)—with blue lines for convolutions, red for downsampling, and green for upsampling. HRNet combines all this multi-scale information for precise image representation. (Right) Disentangled keypoint regression (DEKR). The input to the DEKR module is the feature maps from the last layer of the HRNet backbone. The feature maps are combined at different resolutions to form the first block on the left. Each of these branches learns the representation of a keypoint by applying two adaptive convolutions to a portion of the feature maps generated by the backbone and regressing the 2D offset of that keypoint using a 1x1 convolution. In the figure, this process is shown for three keypoints, with the feature maps divided into three partitions, each assigned to a separate branch. For our purposes the feature maps are divided into 14 partitions, resulting in 14 branches, each responsible for regressing one of the 14 keypoints.

mechanism, depicted in Figure 2 (right). The HRNet backbone processes the input image to produce high-resolution feature maps. These maps are fused and aligned across different scales to form a unified representation, which serves as the input for the DEKR module. This module uses a multi-branch structure to regress precise keypoint positions, where each branch is dedicated to a specific keypoint such as shoulders or elbows.

Within each branch, adaptive convolutions focus on the keypoint region to refine feature processing. Subsequent 1x1 convolutions generate an offset vector predicting the 2D displacement from the center to the keypoint location. The DEKR module first uses these vectors to estimate keypoint coordinates. Then, Non-Maximum Suppression (NMS) is applied in two stages to enhance accuracy: an initial NMS on the center heatmap removes non-maximum points, and a second NMS on the regressed poses eliminates overlapping keypoints, ensuring only the most accurate predictions are retained. The final step combines these vectors with the central positions to determine the coordinates of the keypoints.

During the simulation phase of the monitoring scenario in a NICU, the HRNet+DEKR model processes input from footage from four cameras into a 4-frame grid (2x2). This method replicates a comprehensive support system for monitoring the movements of preterm infants in the NICU and leverages the inherently multi-pose capabilities of the architecture, which, as stated by the original authors [2], can identify up to 30 individuals simultaneously. By using a single HRNet+DEKR model to analyze this grid, instead of deploying four separate models, we harness substantial computational and operational efficiencies. In fact, a single model approach markedly reduces the demand for computational resources such as memory and processing power, streamlines system management, and reduces the complexity of operations. This strategy improves response times, facilitating timely interventions

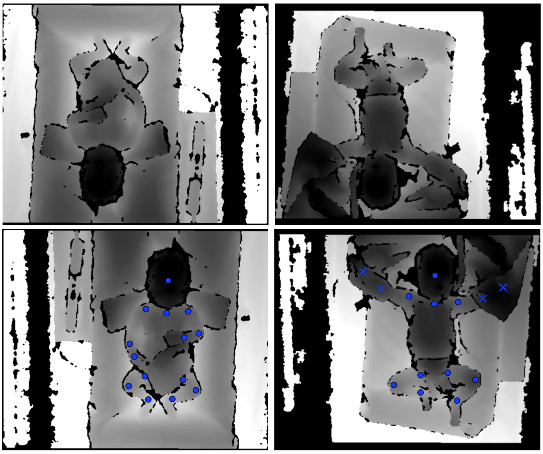

Figure 3: The upper part of the figure displays the original frames from the expanded BabyPose dataset [6]. The lower part of the figure presents the frames after a 180-degree rotation to align with the pre-training on the CrowdPose dataset. The sample image on the left does not show any keypoint occlusion, unlike the image on the right, in which the occlusions are indicated by an "x".

critical in neonatal care. Furthermore, opting to implement this multi-frame integration only during the testing phase allows the system to be flexible, not limiting the number of frames in the grid or the streams from the outset, thereby maintaining scalability.

3 Experimental protocol

3.1 *Dataset*

As introduced in Sec. 1, in this work, we used the expanded BabyPose dataset [6]. This is a collection of depth videos displaying preterm infants in the NICU of the G. Salesi Hospital (Ancona, Italy). For our purposes, we used 18,000 depth frames from 18 preterm infants, extracted from 180-second recordings. The videos were recorded using an Astra Mini S-Orbbec with a frame rate of 30 frames per second and a resolution of 640 × 480.

The original annotation of the BabyPose dataset included the keypoints related to the limbs (i.e., wrists, elbows, shoulders, ankles, knees, hip). However, for this work, the dataset was re-labeled to include 14 keypoints, to align with the keypoint structure used in the pre-trained model from the CrowdPose dataset (more details on pre-training will be provided in Sec. 3.2). In particular, two keypoints were added (i.e., head and sternum). The annotation procedure was carried out under the supervision of our clinical partners using a custom-built online tool, Label studio[2].

The dataset was divided into training, validation, and test sets. Data from two patients, totaling 2,000 frames, were randomly selected for the test set to ensure that these images were never exposed to the network during training or validation. The remaining data from

[2] Label Studio: Open Source Data Labeling - hrefhttps://labelstud.io/

16 patients were used for training and validation, with 12,000 frames allocated for training and 4,000 frames reserved for validation.

Figure 3 (lower part) presents two samples from the dataset and the human-annotated keypoints. As illustrated in the both frames, various objects (such as towels, drug infusion systems, etc.) may be present within the frames and hide parts of the infant's body from camera. Occlusions are shown in the bottom right frame as "x" superimposed to the missing keypoint. For such cases, the annotations take advantage of the visibility parameter in Label Studio, which allows for the notation of an occluded keypoint (in this case right elbow and wrist).

3.2 Training settings

For our study, we used the HRNet+DEKR model. We initialized the HRNet backbone weights with those derived from pre-trained on ImageNet; then the HRNet+DEKR model was pre-trained on the CrowdPose dataset and fine-tuned using the expanded BabyPose dataset. Notably, since HRNet+DEKR was originally trained on upright, real-world images, we rotated the images and annotations from the expanded BabyPose dataset by 180° (Figure 3). This adjustment aligned them with the model's expected input orientation, allowing us to effectively leverage the pre-training.

To train our model, we used the following loss function (l), as described in [2]:

$$l = l_h + \lambda l_p \tag{1}$$

The loss function combines two components: offset loss (l_p) and heatmap estimation loss (l_h).

The *offset loss* (l_p) is designed to refine the precise location of keypoints by measuring the difference between the predicted offsets and the actual offsets (ground truth) of keypoints.

where:

l_p is the *offset loss*:

$$l_p = \sum_{i \in C} \frac{1}{Z_i} \operatorname{smooth}_{\mathrm{L}_1}(\mathbf{o}_i - \mathbf{o}_i^*) \tag{2}$$

with $Z_i = \sqrt{E_i^2 + W_i^2}$ being the person instance size - E_i and W_i are the height and the width of the instance box- C the set of predicted poses matching ground truth annotations, and $\mathbf{o}_i$ and $\mathbf{o}_i^*$ the estimated and ground truth offset vectors, respectively.

The *heatmap estimation loss* (l_h) represents the weighted distance between the heat values predicted by the network and the ground truth heat values.

l_h is the *heatmap estimation loss*:

$$l_h = \|\mathbf{M}^h \odot (\mathbf{H} - \mathbf{H}^*)\|_2^2 + \|\mathbf{M}^c \odot (\mathbf{C} - \mathbf{C}^*)\|_2^2 \tag{3}$$

with H being the predicted keypoint heatmaps, and C the predicted center heatmap. C^* and H^* are the ground truth keypoint and center heatmaps, respectively. The masks M^h and M^c are applied to the keypoint and center heatmaps to focus the loss calculation on relevant areas, enhancing the model's ability to capture the spatial locations of keypoints.

The training process used a learning rate of 0.001 across 100 epochs. Early stopping was implemented to prevent overfitting; training was halted if the validation loss did not improve for 10 consecutive epochs. A batch size of 8 was used, which was the maximum number of images the GPU could handle simultaneously.

To fully leverage the capabilities of the pre-trained model, the depth images were converted into a three-channel format. These images underwent preprocessing to align their distribution with that of the images used during the model's pre-training phase. Specifically, the colour channels of each image were normalized by subtracting the mean and dividing by the standard deviation. Online data augmentation techniques were applied during the training phase, including rotations of ±30°, scaling between 0.75 and 1.5, and translations up to ±40 pixels. In the 2x2 grid test, once the grid was created, the resulting image underwent resizing and pre-processing, following the same procedure as that used for the single-frame input. It is worth noting that the grid's dimension depends on the system's design to balance computational load and the level of detail required for effective monitoring.

3.3 Performance metrics

To evaluate the performance of the pose estimation model, the standard MS-COCO Average Precision (AP) and Average Recall (AR) based on Object Keypoint Similarity (OKS)[3] were used, along with AP.5, AR.5 (AP and AR with OKS threshold of 0.5) and AP.75, AR.75 (AP and AR with OKS threshold of 0.75). To further evaluate the model performance, a newly introduced metric was also used, the Limb Overlap Score (LOS).

The LOS is calculated through a multi-step process that begins by grouping keypoints into four distinct groups: the right arm (i.e., right wrist, elbow, shoulder), left arm (i.e., left wrist, elbow, shoulder), right leg (i.e., right ankle, knee, hip), and left leg (i.e., left ankle, knee, hip). After grouping, the Euclidean distance between the predicted keypoints and their ground truth counterparts is computed for each group, specifically excluding keypoints that are annotated as non-visible (i.e., due to external occlusions, like in Figure 3). The average distance for each group of limbs is then determined. Finally, the LOS is calculated on the basis of these average distances, providing a quantitative measure of the model's ability to predict the limbs' position.

with:

$$\mathrm{LOS} = 1 - \frac{\sum_{i=1}^{3} \sqrt{[(x_{pi}-x_{gi})^2+(y_{pi}-y_{gi})^2]}\,\delta(v_i>0)}{\sum_i \delta(v_i>0)\sqrt{A^2+B^2}} \tag{4}$$

where (x_{p_i}, y_{p_i}) and (x_{g_i}, y_{g_i}) denote the predicted and ground truth coordinates of the *i*-th keypoint of a limb, v_i is the visibility flag, and A and B represent the width and height of the image, respectively. This score normalizes limb detection accuracy by the image diagonal, ensuring that it is independent of image size and adaptable to different image dimensions and aspect ratios.

In contrast to the OKS approach, which assesses keypoint identification on a global scale, LOS assesses the model's accuracy in regressing each limb's position individually (i.e., right arm, left arm, right leg, left leg), excluding non-visible keypoints from its computation. This ensures that the metric reflects only the visible and measurable aspects of the limb's position, providing a more accurate and clinically relevant assessment of the model's performance.

4 Results and discussion

Table 1 presents the model's performance using OKS-based metrics. AP scores across different Intersection over Union (IoU) thresholds—0.5 (AP.5), 0.75 (AP.75), and an average

[3] COCO Keypoint Evaluation

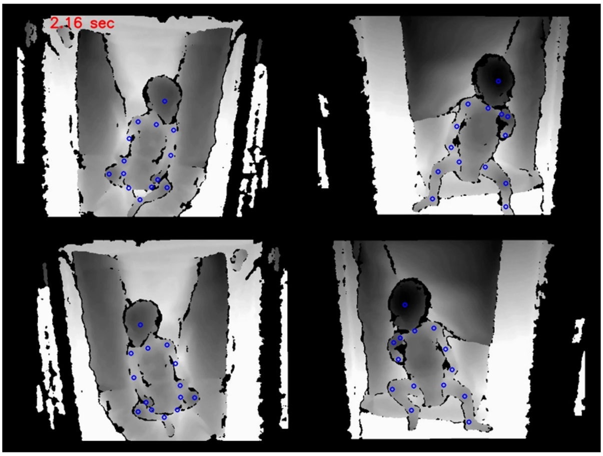

Figure 4: Qualitative results from the High-Resolution Network + Disentangled Keipoint Regression (HRNet+DEKR) on the 2x2 grid

Table 1: Performance of the High-Resolution Network + Disentangled Keypoint Resolution (HRNet+DEKR) in terms of Object Keypoint Similarity (OKS)-based metrics.

AP	**AP.5**	**AP.75**	**AR**	**AR.5**	**AR.75**
0.975	0.989	0.987	0.985	0.998	0.992

across all thresholds (AP)—are exceptionally high, indicating that the model accurately detects keypoints in most cases. In particular, the AP score of 0.975 suggests that the model achieves near-perfect detection in various poses and scenarios. The model performs slightly better at a lower IoU threshold (AP.5 = 0.989), which is typical because the criteria for a correct prediction are less stringent. Similar trends can be seen on the AR.

Table 2 evaluates the HRNet+DEKR model using the LOS for each limb individually and on average. Consistency in performance across all limbs—left arm (0.993), right arm (0.991), left leg (0.992), and right leg (0.992)—highlights the model's ability to equally capture the dynamics of both upper and lower limbs with high precision. The overall average LOS of 0.992 further attests to the uniformity in the accuracy of the model in different body parts.

The high scores in both the OKS-based metrics and LOS suggest that HRNet+DEKR is well-suited for precise pose estimation required in NICUs. The ability to maintain high accuracy across various limbs and under different evaluation metrics (AP, AR, and LOS) suggests that the model can be used reliably in practical applications, where precise detection of movement of the limbs may potentially contribute to monitoring and diagnostic processes.

These results also open avenues for further research into the application of pose estimation technologies in healthcare, especially in contexts where non-invasive monitoring is essential. Future studies could explore the integration of such models with real-time monitoring systems in NICUs, potentially offering a non-contact method to assess the development of preterm infants.

Building upon the results obtained from single-subject data, the HRNet+DEKR model

Table 2: Performance of the High-Resolution Network + Disentangled Keypoint Regression (HRNet+DEKR) in terms of Limb Overlap Score (LOS).

Left Arm	Right Arm	Left Leg	Right Leg	Average
0.993	0.991	0.992	0.992	0.992

was applied to multi-pose estimation. To facilitate this, a dataset was created by arranging multiple patient sequences in a 2x2 grid configuration (as illustrated in Figure 4). As visible from the qualitative results, the proposed HRNet+DEKR is able to accurately estimate the location of the keypoints of interest in each of the 4 patients. This multi-person approach may significantly boost the efficiency of the inference process, meeting a critical need within the ward to lower computational demands and streamline operations. By processing videos simultaneously on a single hardware system, the grid approach reduces the resources needed, thereby facilitating more efficient monitoring and analysis in scenarios involving multiple subjects.

5 Conclusion

This work proposes a DL pipeline based on the HRNet+DEKR model for 2D pose estimation to support the assessment of GMs of preterm infants in the NICU. Its capability to perform multi-pose estimation across a grid of four frames enhances applicability in clinical workflows, where efficiency is crucial. Indeed, by employing a single DL model to monitor multiple infants simultaneously, our approach may both reduce processing time and computational load and – ad a consequence – address the financial constraints hospitals face in installing DL-based monitoring systems.

Future work will quantitatively evaluate the grid approach's performance and assess whether it degrades as grid size increases. This involves systematically varying the dimensions of the grid and measuring the impact on processing speed, accuracy of pose estimation, and use of computational resources. Further, we plan to incorporate machine learning to dynamically optimize grid configurations based on scene complexity and system capabilities.

We are also working to expand the dataset to enhance the performance of the model by encompassing a more extensive range of real-world scenarios, thereby reducing possible biases. Another future step will deal with the integration of a classification approach for GMs.

ACKNOWLEDGMENT

The authors would like to acknowledge the "ITALIAN FUND FOR APPLIED SCIENCES (FISA) grant n°FISA2022-00696" and the National Recovery and Resilience Plan (NRRP), Mission 4, Component 2, Investment 1.1, Call for tender No. 104 published on 2.2.2022 by the Italian Ministry of University and Research (MUR), funded by the European Union – NextGenerationEU– Project Title THAI-MIA – CUP B53D23005000006. Grant Assignment Decree No. 20227HSE83 adopted on May 29, 2023 by the Italian MUR.

References

[1] Christa Einspieler, Arend F Bos, Melissa E Libertus, and Peter B Marschik. The general movement assessment helps us to identify preterm infants at risk for cognitive dysfunction. *Frontiers in Psychology*, 7:406, 2016.

[2] Zigang Geng, Ke Sun, Bin Xiao, Zhaoxiang Zhang, and Jingdong Wang. Bottom-up human pose estimation via disentangled keypoint regression. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14676–14686, 2021.

[3] Xiaohui Gong, Xiao Li, Li Ma, Weilin Tong, Fangyu Shi, Menghan Hu, Xiao-Ping Zhang, Guangjun Yu, and Cheng Yang. Preterm infant general movements assessment via representation learning. *Displays*, 75:102308, 2022.

[4] Daniel G Kyrollos, Anthony Fuller, Kim Greenwood, JoAnn Harrold, and James R Green. Under the cover infant pose estimation using multimodal data. *IEEE Transactions on Instrumentation and Measurement*, 72:1–12, 2023.

[5] Lisa Letzkus, J Vince Pulido, Abiodun Adeyemo, Stephen Baek, and Santina Zanelli. Machine learning approaches to evaluate infants' general movements in the writhing stage—a pilot study. *Scientific Reports*, 14(1):4522, 2024.

[6] Lucia Migliorelli, Sara Moccia, Rocco Pietrini, Virgilio Paolo Carnielli, and Emanuele Frontoni. The babypose dataset. *Data in Brief*, 33:106329, 2020.

[7] Lucia Migliorelli, Emanuele Frontoni, and Sara 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*, 204:117458, 2022.

[8] Lucia Migliorelli, Alessandro Cacciatore, Valeria Ottaviani, Daniele Berardini, Raffaele L Dellaca', Emanuele Frontoni, and Sara Moccia. Twineda: a sustainable deep-learning approach for limb-position estimation in preterm infants' depth images. *Medical & Biological Engineering & Computing*, 61(2):387–397, 2023.

[9] Matteo Moro, Vito Paolo Pastore, Chaira Tacchino, Paola Durand, Isabella Blanchi, Paolo Moretti, Francesca Odone, and Maura Casadio. A markerless pipeline to analyze spontaneous movements of preterm infants. *Computer Methods and Programs in Biomedicine*, 226:107119, 2022.

[10] Haomiao Ni, Yuan Xue, Liya Ma, Qian Zhang, Xiaoye Li, and Sharon X Huang. Semi-supervised body parsing and pose estimation for enhancing infant general movement assessment. *Medical Image Analysis*, 83:102654, 2023.

[11] Kamini Raghuram, Silvia Orlandi, Paige Church, Tom Chau, Elizabeth Uleryk, Petros Pechlivanoglou, and Vibhuti 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):637–648, 2021.

[12] Ameur Soualmi, Christophe Ducottet, Hugues Patural, Antoine Giraud, and Olivier Alata. A 3d pose estimation framework for preterm infants hospitalized in the neonatal unit. *Multimedia Tools and Applications*, 83(8):24383–24400, 2024.

[13] Hélène Turpin, Sébastien Urben, François Ansermet, Ayala Borghini, Micah M Murray, and Carole Müller-Nix. The interplay between prematurity, maternal stress and children's intelligence quotient at age 11: A longitudinal study. *Scientific Reports*, 9 (1):450, 2019.

[14] Ce Zheng, Wenhan Wu, Chen Chen, Taojiannan Yang, Sijie Zhu, Ju Shen, Nasser Kehtarnavaz, and Mubarak Shah. Deep learning-based human pose estimation: A survey. *ACM Computing Surveys*, 56(1):1–37, 2023.