Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

NICUPose: a robust and efficient end-to-end pose estimation network in neonatal intensive care unit (NICU)

Di Huang
Washington University in St. Louis
Alec Gleason
Albert Einstein College of Medicine
Richter Felix
Icahn School of Medicine at Mount Sinai
John Zempel
Washington University School of Medicine
Philip Payne
Washington University School of Medicine
Yixin Chen
Washington University in St. Louis
F. Cole
Washington University School of Medicine
Zachary Vesoulis
Washington University School of Medicine
Fuhai Li
fuhai.li@wustl.edu

Washington University School of Medicine

Article

Keywords:

Posted Date: November 11th, 2024

DOI: https://doi.org/10.21203/rs.3.rs-5349177/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

NICUPose: a robust and efficient end-to-end pose estimation network in neonatal intensive care unit (NICU)

Di Huang[1,2], Alec Gleason[6], Richter Felix[3], John M. Zempel[4,5], Philip R.O. Payne[1], Yixin Chen[2], F. Sessions Cole[5], Zachary A. Vesoulis[4,5#], Fuhai Li[1,5#]

[1]Institute for Informatics (I2), Washington University School of Medicine, St Louis, MO, USA
[2]Department of computer science and engineering, Washington University in St Louis, St Louis, MO, USA
[3]Department of Pediatrics, Icahn School of Medicine at Mount Sinai, New York, NY, USA
[4]Department of Neurology, Washington University School of Medicine, St Louis, MO, USA
[5]Department of Pediatrics, Washington University School of Medicine, St Louis, MO, USA
[6]Einstein College of Medicine, Bronx, New York, USA

Correspondence: vesoulis_z@wustl.edu; Fuhai.Li@wustl.edu

Abstract – In the neonatal intensive care unit (NICU), long-term continuous video-electroencephalography (vEEG) monitoring is used for diagnosis, prognosis, and monitoring of neonatal diseases including seizures and other cerebral injury. vEEG includes both an electrographic component as well as synchronized video, which must be manually reviewed for interpretation. This is a time-consuming, labor-intensive, and expensive process which often lags behind real time. While electrographic data are the principal basis for diagnosis, video provides corroborating evidence both for artifact detection (e.g., provider examining patient) and physical manifestation of seizures. This is the first study to focus on neonatal pose estimation in this context. The model's performance was systematically evaluated using two independent vEEG datasets from the Washington University School of Medicine and the Icahn School of Medicine at Mount

Sinai, in addition to publicly available infant datasets. NICUPose demonstrated superior performance and efficiency compared to generic baseline models for pose estimation, highlighting its potential as a valuable clinical tool.

Introduction

The movements of neonates in the Neonatal Intensive Care Unit (NICU) provide critical insights into their neurological status, potential injuries, and future developmental risks. Abnormal movement patterns, such as those linked to cerebral palsy or seizure disorders, can serve as early indicators of neurological dysfunction [1][2][3][4][5][6]. Early detection of these patterns is crucial for timely intervention, potentially enabling treatments like seizure management or early developmental therapies that improve long-term outcomes.

To facilitate early detection, continuous video-electroencephalography (vEEG) has become an essential clinical tool for diagnosing and monitoring neurological conditions in NICU neonates at risk of injury, particularly seizures [7]. Standard neonatal vEEG combines electrographic data from electrodes placed on the scalp with synchronized video recordings, offering a comprehensive view of the infant's neurological state. As the gold standard for neonatal seizure detection, vEEG can provide reliable prognostic information within just 48 hours [8]. However, current clinical practice largely relies on manual, subjective interpretation of this data. The video component, while crucial, is underutilized by automated systems and remains essential for human reviewers who use it to corroborate seizure activity, differentiate between true seizures and physical

handling errors, and perform basic virtual physical assessments [1][2][4][5][6]. While algorithms exist for the automated detection of seizures [9][10][11], their reliance on electrographic data alone limits their effectiveness. Incorporating pose estimation specifically designed for the NICU environment could address this limitation by mirroring human interpretation of both electrographic and video data. This could streamline the currently labor-intensive process, enhancing accuracy and efficiency in diagnosing neonatal neurological conditions.

Despite the potential of deep learning-based pose estimation for automatic video-based movement assessment, significant challenges remain in adapting these techniques to the NICU environment. Pose estimation generally involves detecting and linking key body points—such as the ankles and wrists—to represent an individual's skeletal structure. When applied longitudinally, serial pose tracking can offer a detailed assessment of movement patterns. While the field of pose estimation has advanced considerably with computer vision technologies, much of the progress has centered around adult pose estimation [13][14][15][16][17][18][19][20], with more recent adaptations targeting non-hospitalized newborns [12][21] and, to a limited extent, hospitalized newborns [22][23][24][25]. However, no existing studies have specifically tailored pose estimation networks for neonates in the NICU who are undergoing vEEG monitoring and are at heightened risk of brain injury. This gap underscores the need for a deeper investigation into the unique challenges posed by the NICU environment.

While previous studies have introduced infant pose datasets and corresponding pose estimation networks, they have not adequately addressed the unique challenges faced by NICU newborns undergoing vEEG monitoring. For example, studies utilized infant images sourced from YouTube or Google Image searches [21], while others focused on RGB video frames of hospitalized newborns [22][23] or employed depth imaging techniques [24][25]. These datasets often depict infants in controlled settings, typically centered in the frame, lying on simple backgrounds such as blankets, and wearing minimal clothing, like diapers. In contrast, video footage of NICU newborns presents a far more complex scenario. Factors such as variable camera angles, neonatal positioning, body occlusion from medical equipment, fluctuating lighting conditions, and ongoing clinical interventions differ significantly between recordings, and sometimes even within the same video. These challenges create a highly dynamic and unpredictable environment for pose estimation, which is explored in detail in the following sections.

Additionally, prior studies have often overlooked the critical need for robustness and efficiency in NICU applications. Robustness refers to the network's ability to handle the highly variable conditions of the NICU environment, while efficiency refers to the speed with which newborn poses can be detected in video frames—both of which are essential for timely clinical interventions. Current infant pose estimation networks, whether employing top-down [21][22] or bottom-up [23][24][25] approaches, are constrained by their suboptimal two-stage design, limiting their suitability for use in the NICU.

In this study, we address two critical gaps in NICU newborn pose estimation: the lack of representative NICU datasets and the absence of an end-to-end pose estimation network tailored for NICU neonates undergoing vEEG monitoring. To overcome these challenges, we introduce NICUPose, a novel single-stage, end-to-end network specifically designed for NICU settings. Our contributions include: (1) the development of two independent, annotated NICU pose datasets from multicenter vEEG video recordings, capturing a wide range of neonatal poses and clinical conditions; (2) a quantitative visual analysis that highlights significant differences between NICU and non-ICU infant datasets; (3) the design of an end-to-end vision transformer-based network that achieves superior detection accuracy and speed compared to four state-of-the-art models; and (4) comprehensive evaluations demonstrating the robustness and efficiency of NICUPose on real-world NICU data, where it achieved the highest mean Average Precision (mAP) and fastest inference speed, significantly outperforming existing methods.

Results

Comparative analysis of visual features in NICU newborns and other infant contexts

To illustrate the unique and complex visual characteristics of NICU newborns in vEEG monitoring, we conducted a comparative analysis to differentiate these features from those observed in other infant contexts. This comparison included NICU newborn in vEEG (WashU-NICU) [40], hospitalized ICU newborns represented in depth images (BabyPose) [26], non-hospitalized infants (SyRIP) [21], and synthetic hospitalized

infants (MINI-RGBD) [27], each of which represents public research datasets commonly used as training sources in recent literature.

We sampled 500 images at a resolution of 640x480 from each dataset to ensure a diverse representation of infants. We employed DINO [29], a self-supervised learning approach, to train a Vision Transformer (ViT) [35], a foundational vision network, on four datasets: WashU-NICU, SyRIP, MINI-RGBD, and BabyPose, with each dataset treated as a distinct class. The training objective of this self-supervised method was to enable the network to learn the visual representations of the input data. We then used the pretrained network for clustering the class features by projecting them into a two-dimensional space, allowing for the quantification of visual features in the computer vision representation domain.

We visualized the extracted visual features using the attention heatmap from the pretrained ViT, shown in **Figure 1a**. The first attention heatmap highlights the infant subject, while the second highlights the background, effectively demonstrating the feature extraction capabilities of this self-supervised method. Next, we applied Uniform Manifold Approximation and Projection (U-MAP) [36] to project the extracted features from each class into a 2-D feature space, shown in **Figure 1b**. Distinct patterns emerged: real infant datasets, including WashU-NICU and SyRIP, naturally formed discernible and distinct

clusters, while synthetic and depth-image datasets, such as MINI-RGBD and BabyPose, exhibited more scattered distributions with less unified clustering.

We further identified that contextual elements significantly affect visual representation in computer vision. Both real RGB datasets feature images with diverse infant instances, backgrounds, and body poses, while synthetic datasets lack such variation. Both the WashU-NICU and SyRIP datasets contain diverse and colorful images of infants, featuring a variety of poses, backgrounds, and clothing. This variability allows the pretrained network to learn a universal representation of infants, resulting in unified clusters. In contrast, MINI-RGBD and BabyPose datasets, with limited pose variation and constant backgrounds, display scattered clustering due to their invariant nature.

This discrepancy in visual features underscores a critical insight: pose estimation networks trained on datasets that differ from the RGB format used for NICU newborns exhibit weaker visual understanding of this specific group. As a result, these networks may produce inaccurate pose estimations when applied to NICU newborn data.

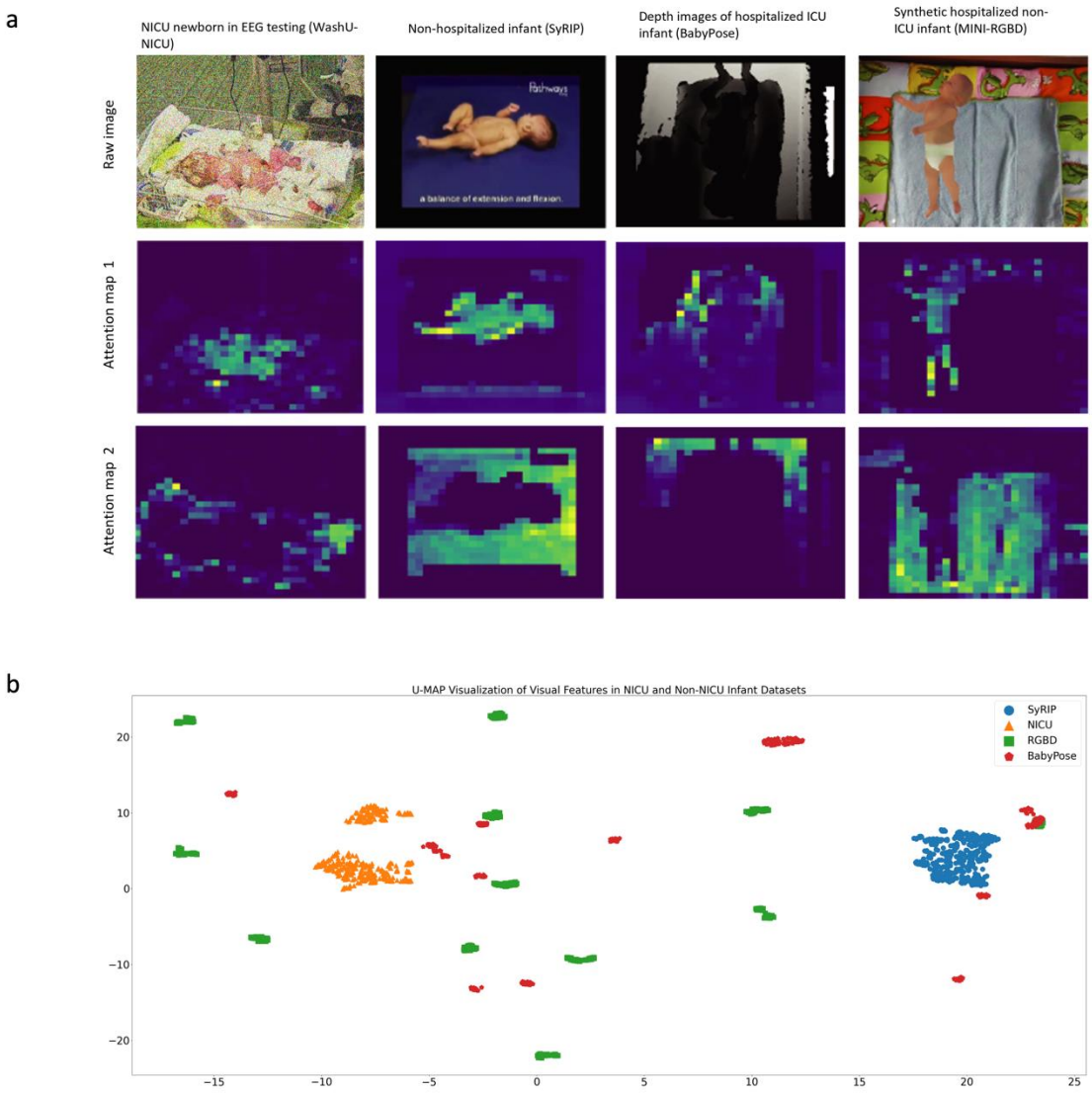


Figure 1| Visual feature analysis among NICU newborn in vEEG (WashU-NICU), non-hospitalized infant (SyRIP:), depth-image of ICU hospitalized newborn (BabyPose), synthetic hospitalized infant (MINI-RGBD). a, Visualization of attention heatmap of pretrain ViT on NICU newborn and other infant dataset**.** The first row shows the self-attention heatmap shows the infant instance. The second row shows the attention heatmap of the background. **b**, U-MAP visualization of the visual features in 2-D. The orange dot denotes the projected visual features of our WashU-NICU newborn, and the blue dot denotes that of SyRIP, the green square denotes that of MINI-RGBD, the red pentagon denotes that of the Baby Pose

Unique challenges in NICU pose estimation

We identified three primary challenges unique to the NICU setting: **occlusion**, **color inconsistency**, and **contextual confusion**, as illustrated in **Figure 2**. These

challenges are rarely present in non-ICU infant datasets but are pervasive in NICU newborns during EEG monitoring, significantly complicating pose estimation.

Occlusion occurs due to the newborn's various lying positions and the extensive clinical equipment attached to their bodies. Newborns frequently shift positions during prolonged EEG monitoring, and camera displacement over time further complicates the situation. For instance, in **Figure 2 (top-left)**, a newborn lying in a lateral position obstructs the visibility of certain body parts. Similarly, in **Figure 2 (top-right)**, medical devices such as electrode caps, blankets, and respiratory support systems further contribute to occlusion, leading to incomplete visibility of the newborn’s body. As a result, pose estimation models must be designed to accurately detect visible keypoints, even when parts of the body are occluded by equipment or the newborn's position.

Color inconsistency arises from the varying illumination conditions typical of the NICU environment. Overhead lights are turned on and off during examinations and routine patient care, and lighting can vary depending on the time of day and the specific needs of the infant. Some patient rooms have windows that allow natural light, while others are equipped with phototherapy lamps to treat conditions such as hyperbilirubinemia. Some NICU rooms are kept dark to minimize stimulation in critically ill newborns, requiring the use of infrared lamps mounted on EEG carts for continued illumination. When infrared lighting is in use, images are captured in grayscale, as shown in **Figure 2 (bottom-left)**. This variability in color format, contrast, and lighting conditions presents a significant

challenge for pose estimation models, which must be robust enough to handle these fluctuations.

Contextual confusion arises from clinician interventions, typically occurring every three hours or more frequently depending on the newborn’s condition. These interventions introduce the presence of clinician body parts, such as hands and wrists, into the frame, complicating the identification of the newborn’s keypoints. An example of this is shown in **Figure 2 (bottom-right)**, where a clinician's wrists are visible. These occurrences introduce visual noise into the pose estimation task, as the model must differentiate between the newborn's keypoints and those of the clinician, which are often similar in shape and position.



Figure 2 | Visual challenges of NICU newborn in vEEG. The images are sampled from WashU-NICU dataset with added noise. Newborn lying in lateral position (top-left), body occlusion (top-right), lights-off environment (bottom-left), and clinician intervention (bottom-left).

NICUPose network architecture

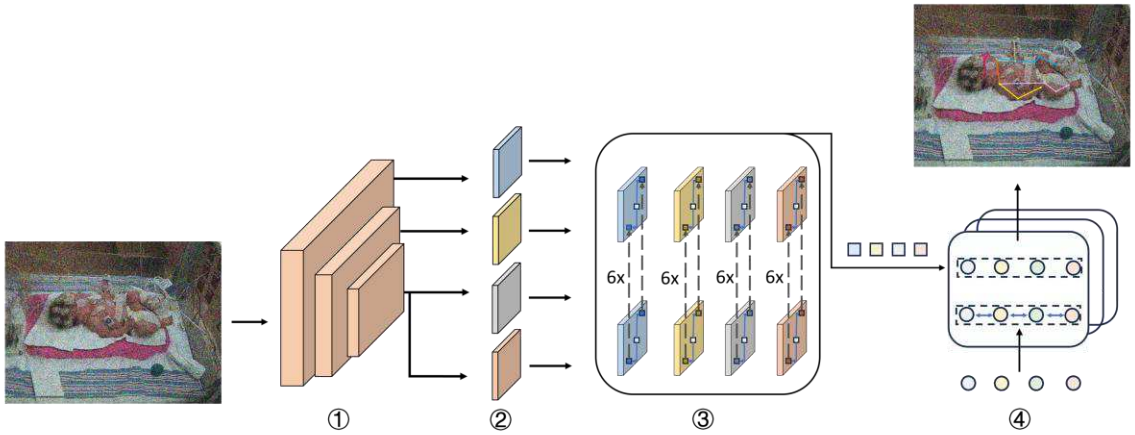
Recognizing the necessity of using real NICU newborn vEEG RGB data for network training, we propose NICUPose, a single-stage, end-to-end transformer specifically designed for pose estimation in NICU newborns. NICUPose takes an input image of a NICU newborn and directly generates the poses of the newborn. NICUPose consists of three components, as shown in **Figure 3a**: a backbone network, a deformable encoder, and a deformable decoder. The backbone, labeled ① in **Figure 3a**, is a pretrained vision foundation network that extracts multi-scale feature maps from the input image, labeled ② in **Figure 3a**. By leveraging these multi-scale features, the network is robust to scale variations in the input data, capturing both fine-grained and coarse-grained features at different scales. Scale variance is common in NICU settings, as the camera's position may shift over time, causing the newborn to appear at varying scales in the video. Incorporating multi-scale features helps mitigate the impact of these scale variations, improving pose estimation for NICU newborns.

We adopt the deformable encoder introduced in previous work [37], which is known for its computationally efficient deformable attention mechanism. The deformable encoder, labeled ③ in **Figure 3b**, processes the features extracted by the backbone to generate a set of encoded keypoint queries. It captures spatial and contextual representations of the keypoints, enabling the network to identify potential keypoints and refine them for pose

estimation. We chose this deformable encoder design due to its deformable attention unit, which allows the transformer-based network to fully exploit the multi-scale feature map while maintaining linear-time computation. This results in higher accuracy and faster training and inference speeds [16][37][38] compared to traditional scaled-dot product dense attention mechanisms [39].

The deformable decoder, labeled ④ in **Figure 3c**, also uses the deformable attention mechanism to reason about the keypoint features generated by the encoder and estimate the final pose. It takes randomly generated pose queries and interacts with the keypoint features to refine these queries into the final predicted poses. This deformable encoder-decoder architecture allows us to create an end-to-end pose estimation network optimized for NICU newborns.

a



Multi-scale feature map

Multi-scale deformable self-attention

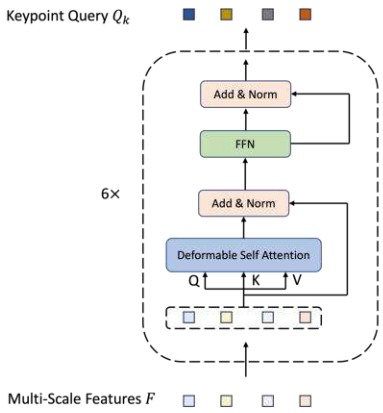
Multi-scale deformable cross-attention

Scaled dot-product self-attention

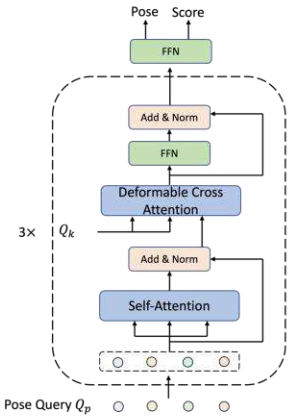
Keypoint query (Encoder query)

Pose query (Decoder query)

b



c



d

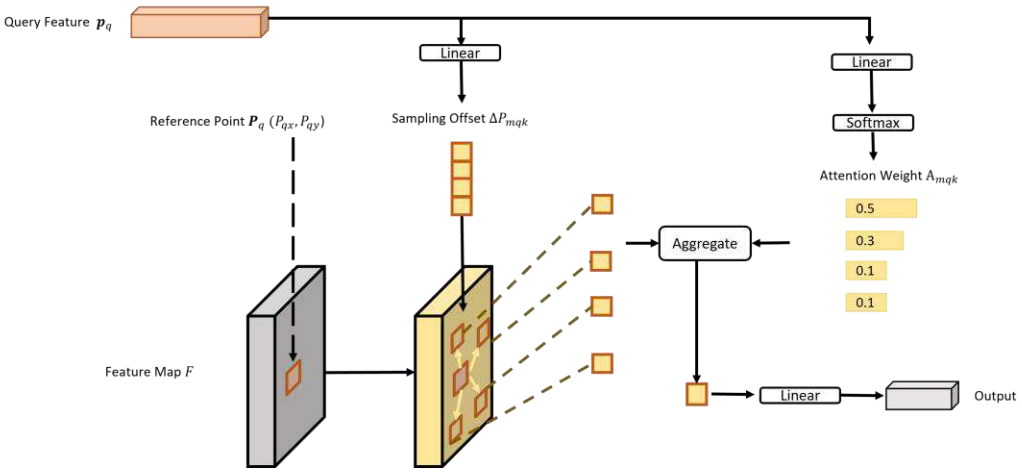


Figure 2 | **NICUPose Network Architecture. a**, Visualization of NICUPose network. The input to the network is an image. This image is first processed by the backbone network, labeled with ①, to generate feature maps at four different scales, labeled with ②. We mark the feature maps in light blue, yellow, light gray, and beige. Each feature map is fed into the deformable encoder, labeled with ③, to generate keypoint queries (features). The encoder contains six identical encoder layers and processes feature maps at each scale in parallel within each layer. The processed keypoint features from each layer are fed into the next layer iteratively. During the decoding phase, the three-layer deformable decoder, labeled with ④, intakes randomly generated pose queries. The pose queries first go through a scaled dot-product self-attention and then through a deformable cross-attention with the refined keypoint queries to formalize the pose. The output of the last decoder layer is the detected poses, which are in-painted on the original image. **b**, The architecture of the deformable encoder, corresponding to ③, consists of six sequential layers. Each layer includes a deformable self-attention module, residual connection, layer normalization (Add & Norm), and feed-forward network. The encoder takes multi-scale features as input and refines them to generate keypoint queries (features) as output. **c**, The architecture of the deformable decoder, corresponding to ④, consists of three decoder layers. Each layer includes a self-attention module, residual connection, layer normalization (Add & Norm) module, deformable cross-attention module, and feed-forward module. The input to the decoder is a randomly generated pose query, and the output is the predicted poses. **d**, The visualization of single-scale deformable attention. For a given feature query p_q in a single-scale feature map F, it attends to four keys (also known as "queries") in F for sparse attention computation. The locations of the four keys relative to the feature maps are updated through network training, which saves computational complexity. The updated keys allow p_q to attend only to the most important and informative keys.

Pose evaluation

We used MS-COCO [42], SyRIP, Mount Sinai NICU [34], and WashU-NICU as benchmark datasets for pose estimation, as they all follow the same annotation format. MINI-RGBD and BabyPose were not included in the pose estimation evaluation due to differences in annotation formats. Specifically, COCO-train2017, consisting of 64,115

human images, and COCO-val2017, consisting of 2,693 human images, were used for general human pose training and evaluation. SyRIP-val500, consisting of 500 non-hospitalized infant images was used for zero-shot infant pose evaluation. WashU-NICU-train350, consisting of 350 NICU newborn images acquired from St. Louis Children's Hospital, and WashU-NICU-val150, consisting of 150 newborn images, were used for NICU newborn pose fine-tuning and evaluation. Mount-Sinai-v200, consisting of 200 NICU newborn images acquired from Mount Sinai Hospital in New York City, was used for zero-shot evaluation on NICU newborns. Given the large-scale of training data in COCO, neither the infant dataset nor the NICU newborn datasets are comparable in terms of the quantity of training samples. Hence, all networks are initially trained on the COCO-train2017 dataset to ensure networks capturing general human skeleton structure. We provided an overview of the datasets in **Table 1**, listing the subject types, number of training and validation images.

Dataset	Subject	Train set (number of images)	Validation set (number of images)
MS-COCO	Human	64115	2693
SyRIP	Non-hospitalized infant	NA	500
Mount Sinai NICU (ours)	Critically ill ICU newborn in EEG testing	NA	200
WashU-NICU (ours)	Critically ill ICU newborn in EEG testing	350	150

Table 1| Dataset overview of pose estimation benchmark dataset. The subject, number of images in training set, and validation set is provided.

In our experiment, we compared NICUPose with four baseline networks, two top-down approaches, HRNet [17], and HRFormer [19], and two bottom-down approaches, HrHRNet [19] and DKER [20], across five experiments (see **Table 2** to **Table 6**). Our goal is to provide comparison between the two-stage methods and our single-stage method

based on two key factors in pose estimation: detection accuracy measured in mAP (the higher the better), detection speed measured in $second/image$ (the smaller the better).

Pose estimation on human

We evaluated the performance of all networks on COCO-val2017, a standard, large-scale human pose estimation benchmark, to assess their capabilities in general human pose estimation. As shown in **Table 2**, the top-down method, HRNet achieved 74.4 mAP, HRFormer achieved 75.6 mAP, the bottom-up method, HrHRNet achieved 67.1 mAP, DEKR achieved 68.0 mAP, and our end-to-end method, NICUPose achieved 70.5 mAP. NICUPose surpassed the bottom-up while behind the top-down methods in this large-scale human pose evaluation, indicating a sufficient capability of our end-to-end structure for a general human pose estimation. This result also reveals a notable performance difference between top-down and bottom-up methods.

Pose estimation on non-hospitalized infant

Since all networks have the capability for human pose estimation, we evaluated them on the non-hospitalized infant dataset, SyRIP-val500, to determine whether the domain shift from adult to infant impacts performance. Interestingly, the domain shift increases detection accuracy, with all networks achieving over 96 mAP, an approximate 20 mAP increase, as shown in **Table 3**. Our NICUPose network achieves 97.8 mAP, the best performance in non-hospitalized infant pose estimation. This performance boost is notable given the differences in body size, clothing, and activity levels between adults and infants [21].

A closer look at the COCO training dataset revealed the inclusion of at least 410 images featuring children. Additionally, the SyRIP dataset primarily consists of individual infants occupying a large portion of the image, with simplified backgrounds that make infant segmentation and pose estimation easier. This reduces the complexity typically found in general human pose datasets, which feature a wider variety of subjects and environments. As a result, without any fine-tuning on the non-hospitalized infant dataset, all networks achieved near-perfect detection accuracy. This suggests that infant pose estimation is not particularly challenging, as training on general human datasets equips networks to perform infant pose estimation effectively.

Pose estimation on NICU newborns

We first evaluated the zero-shot performance of networks that were previously trained on the COCO dataset, applied to WashU-NICU-val150. As discussed earlier, networks trained on the COCO dataset develop an understanding of poses for adults and non-hospitalized infants. This zero-shot evaluation assessed the feasibility of applying networks with non-hospitalized infant pose knowledge to NICU newborns. As shown in **Table 4**, all networks failed to detect poses in NICU newborn subjects, with the highest score being only 2.5 mAP. This outcome supports our observations regarding the visual feature disparities between non-hospitalized infants and hospitalized ICU newborns. We conclude that the visual features in these two datasets differ substantially, preventing the networks from generalizing to NICU newborns when trained on non-hospitalized infant data.

Next, we fine-tuned all networks on WashU-NICU-train350 dataset to assess the effectiveness of fine-tuning for NICU newborn pose estimation and to identify performance distinctions among the networks. As shown in **Table 5**, all networks showed significant accuracy improvements: HRNet achieved 75.6 mAP, HRFormer achieved 73.6 mAP HrHRNet achieved 80.8 mAP, DEKR achieved 71.4 mAP, and NICUPose achieved 86.5 mAP. To ensure the effectiveness of top-down networks, we separately finetuned the detection network required for this method, which resulted in 80.8 mAP. Unlike in general human pose estimation (as shown in **Table 2),** our end-to-end method significantly outperformed the top-down and bottom-up methods, with an average improvement of 10 mAP. The performance gap between the top-down and bottom-up methods was smaller compared to human pose estimation. Interestingly, HrHRNet, which showed the lowest accuracy in human pose estimation, outperformed other two-stage methods in NICU newborn pose estimation. To further assess the effectiveness and robustness of the networks, we conducted a zero-shot evaluation on an external NICU dataset, Mount Sinai NICU-val200.

As shown in **Table 6**, comparing to the finetuning result on WashU-NICU, NICUPose achieved the highest 71.4 mAP though this was a decrease of 15.1 mAP from the fine-tuning results on WashU-NICU. In contrast, the two-stage networks experienced more substantial performance drops. For the top-down method, HRNet dropped by 17.8 mAP to 57.8 mAP and HRFormer dropped by 15.7 mAP to 57.9 mAP. For the bottom-up method, HrHRNet showed a significant drop from 72.8 mAP to 8.0 mAP and DEKR

dropped from 31.1 mAP to 40.3 mAP. This performance drop is expected, as the NICU environment at Mount Sinai differs from that of WashU-NICU, introducing new variables. However, NICUPose's relatively smaller performance drop demonstrates its robustness and adaptability across different NICU environments, whereas the sharp decline in performance for the two-stage networks highlights their difficulty in adapting to new environments without additional fine-tuning.

We also evaluated the detection speed of the finetuned networks, as timely results are equally important. As shown in **Table 5**, NICUPose achieved the fastest detection speed at 156 ms per image on WashU-NICU, which is nearly 1.5 times faster than the second fastest method, HRNet. Specifically, the top-down methods HRNet and HRFormer achieved 234 ms and 338 ms per image, respectively. The bottom-up methods HrHRNet and DEKR achieved 490 ms and 314 ms per image, respectively. The efficiency of our end-to-end network, powered by its deformable attention mechanism, makes it more computationally efficient than the two-stage methods.

Network Analysis

We concluded that the network architecture design—whether top-down, bottom-up, or our end-to-end approach—plays a significant role in NICU newborn pose estimation.

In terms of robustness, NICUPose achieves the best detection accuracy in the two NICU newborn dataset and experienced the least performance drop in zero-shot evaluations

across different these two NICU datasets. The top-down method outperformed the bottom-up method in both detection accuracy and robustness.

The top-down method is characterized by its two-stage design, consisting of a human detector followed by a pose estimator. Typically, the upstream detection network isolates the human region, simplifying the downstream pose estimation task by focusing on individual keypoints within a narrowed detection area. While this approach reduces the task of pose estimator, it limits the estimator's ability to handle challenging conditions independently—such as those commonly found in NICUs, where medical equipment causes occlusions, lighting is inconsistent, and the medical environment is visually complex. Moreover, inaccuracies in the initial detection stage—such as failure to accurately isolate the newborn in the noisy and complex NICU environment—can cascade through the network, leading to reduced precision in pose estimation. This lack of robust design is particularly evident in fine-tuning on NICU newborns, where the top-down method underperforms compared to its performance in general human pose estimation.

The bottom-up method, meanwhile, falls behind the top-down method due to its approach of detecting individual keypoints and subsequently grouping them into distinct poses. This grouping process is heuristic and prone to errors, especially when keypoints are close together or occluded, leading to incorrect associations and reduced pose estimation accuracy. Bottom-up methods also often rely heavily on local image context for keypoint detection, missing the global context needed to fully understand the pose. This limitation

leads to further inaccuracies, particularly in complex environments such as NICUs. In these settings, frequent occlusions and “noisy” keypoints introduced by clinicians exacerbate the challenges. This explains why the bottom-up method falls behind the top-down method in detection accuracy and zero-shot evaluation. The top-down method mitigates some of this confusion by first extracting the newborn subject, which helps reduce deviations in keypoint detection by narrowing the focus to the isolated subject.

Our network, NICUPose, benefits from its end-to-end structure, which avoids the distinct stages of person detection required in the top-down approach and the keypoint grouping algorithm required in the bottom-up approach. NICUPose directly generates the final poses from the input image in a single pass. This is achieved by the encoder-decoder design in the end-to-end transformer-based network, where the encoder generates keypoint features and directly interacts with the decoder to transform these features into poses. This approach allows direct optimization of network components (encoder and decoder) towards the target, unlike the two-stage methods, where the optimization for each component is independent. Even in challenging environments like NICUs, this unified optimization enables NICUPose to adapt to obstacles simultaneously. Moreover, NICUPose adapts to multi-scale input features, allowing it to capture both local and global context, which leads to a deeper understanding of the image. As a result, NICUPose achieves robust performance in NICU newborn pose estimation, with minimal performance drop in zero-shot evaluations across different NICU datasets. Its components are already well adapted to the complexities of NICU environments and the unique anatomical characteristics of newborns.

In terms of efficiency, our end-to-end approach achieved the fastest detection speed in NICU newborn pose estimation, with the top-down method being the second fastest, and the bottom-up method being the slowest.

The two-stage design inherently introduces computational overhead, as seen in both the top-down and bottom-up methods. In the NICU scenario, where typically only one newborn is present in the image, the top-down method is more efficient than the bottom-up method, as it scales with the number of newborn instances. Ideally, it only needs to detect and estimate the pose of a single newborn. However, the bottom-up method is less scalable, as it must detect all potential keypoints across the entire image and then group them into poses, regardless of the number of newborns, which slows detection speed.

Our end-to-end network, NICUPose, avoids the heavy, independent computational components inherent in two-stage methods. The encoder in NICUPose generates a fixed number of keypoint features rather than searching for potential human regions or keypoints, making its scalability dependent on the input image size, not the number of human instances. The core computational unit in our network is a lightweight, sparse computation unit, ensuring fast training and inference speeds. As a result, NICUPose is more stable and faster in detection, while the top-down method's efficiency decreases with more human subjects, and the bottom-up method remains inherently slower.

Dataset (Train/Eval)	COCO-train2017 / COCO-val2017					
Approach	**Method**	**mAP**	**mAP^{50}**	**mAP^{75}**	**mAP^{M}**	**mAP^{L}**
Top-down	HRNet	74.4	91.1	82.6	72.1	82.2
	HRFormer	75.6	90.6	82.6	70.0	78.2
Bottom-up	HrHRNet	67.1	87.1	74.9	63.6	77.2
	DEKR	68.0	88.3	77.7	67.3	78.1
End-to-end	NICUPose (ours)	70.5	90.4	79.0	65.3	80.5

Table 2 | Pose evaluation on human pose. Networks are trained on trained on the COCO-train2017 and evaluated on MS-COCO-val2017.

Dataset (Train/Eval)	COCO-train2017 / SyRIP-val500					
Approach	**Method**	**mAP**	**mAP^{50}**	**mAP^{75}**	**mAP^{M}**	**mAP^{L}**
Top-down	HRNet	97.7	98.9	98.8	98.8	100
	HRFormer	96.4	99.0	98.7	100	96.4
Bottom-up	HrHRNet	96.2	99.0	98.7	100	96.4
	DEKR	96.2	98.0	97.5	100	96.5
End-to-end	NICUPose (ours)	97.8	98.7	98.5	100	97.8

Table 3 | Zero-shot pose evaluation on non-hospitalized infant. Networks are trained on trained on the COCO-train2017 and evaluated on SyRIP-val500.

Dataset (Train/Eval)	COCO-train2017 / WashU-NICU-val150					
Approach	**Method**	**mAP**	**mAP^{50}**	**mAP^{75}**	**mAP^{M}**	**mAP^{L}**
Top-down	HRNet	2.29	15.3	0	7.5	3.00
	HRFormer	1.50	12.4	0	0	1.60
Bottom-up	HrHRNet	2.40	11.3	0	0	2.50
	DEKR	0	0	0	0	0
End-to-end	NICUPose (ours)	2.50	12.3	0	0.50	2.60

Table 4 | Zero-shot pose evaluation on hospitalized ICU newborn. Networks are trained on trained on the COCO-train2017 and evaluated on WashU-NICU-val150.

Dataset (Finetune/Eval)	WashU-NICU-train350 / WashU-NICU-val150						
Approach	**Method**	**mAP**	**mAP^{50}**	**mAP^{75}**	**mAP^{M}**	**mAP^{L}**	**Time [ms]**
Top-down	HRNet	75.6	100	91.7	82.6	74.8	234
	HRFormer	73.6	90.6	82.6	70.0	78.2	338
Bottom-up	HrHRNet	80.8	100	97.7	85.0	84.2	490
	DEKR	71.4	88.3	77.7	67.3	78.1	314
End-to-end	NICUPose (ours)	86.5	100	97.7	85.1	80.4	156

Table 5 | Fine-tuned pose evaluation on hospitalized ICU newborn. Networks are trained on trained on the COCO-train2017, fine-tuned on WashU-NICU-train350 and evaluated on WashU-NICU-val150.

Dataset (Finetune/Eval)	WashU-NICU-train350 / Mount-Sinai-val200					
Approach	**Method**	**mAP**	**mAP^{50}**	**mAP^{75}**	**mAP^{M}**	**mAP^{L}**
Top-down	HRNet	57.8	74.0	61.8	100	58.6
	HRFormer	57.9	72.7	59.4	100	58.1
Bottom-up	HrHRNet	8.00	19.9	5.50	100	8.20
	DEKR	40.3	49.7	42.8	100	44.9
End-to-end	NICUPose (ours)	71.4	79.6	76.5	100	74.7

Table 6 | Zero-shot pose evaluation on hospitalized ICU newborn. Networks are trained on trained on the COCO-train2017, fine-tuned on WashU-NICU-train350 and evaluated on Mount-Sinai-val200.

Discussion

In this study, we introduced NICUPose, a novel pose estimation network specifically designed for NICU newborns in EEG studies. We identified three key challenges that characterize NICU newborns: occlusions, color inconsistency, and contextual confusion. These challenges are unique and common in the NICU environment but are less prevalent in datasets used in previous infant pose estimation studies, which often involve hospitalized or non-hospitalized infants under more controlled conditions.

We also evaluated existing top-down and bottom-up pose estimation networks on both infant pose datasets and NICU newborn datasets. These evaluations revealed inherent limitations in their two-stage network designs, which reduced robustness, and slowed detection efficiency. To address these issues, we enforced NICUPose as an end-to-end, single-stage pose estimation network. Our goal was to eliminate independent components in the network, allowing it to be optimized in a single pass, thus improving both robustness and efficiency. NICUPose utilizes an encoder-decoder architecture designed to generate final poses directly from the input.

We evaluated NICUPose against existing methods from a comprehensive perspective, including detection accuracy, detection speed, and robustness. These factors are less discussed in infant pose estimation studies, which mainly focus on evaluating pose detection accuracy within a single infant dataset. Our experiments on general human pose estimation demonstrated that NICUPose is comparable to existing state-of-the-art networks. In the evaluation of NICU newborn poses, our network achieved the highest detection accuracy, the fastest detection speed. Moreover, in a zero-shot evaluation on an external NICU newborn pose dataset, our network experienced the least performance drop, indicating its robustness.

For future work, we intend to use our network as a semi-automatic labeling tool to annotate more NICU newborn videos available in St. Louis Children's Hospital. This will allow us to expand our dataset and further improve our network as more data becomes available. Additionally, we acknowledge that detecting NICU newborn poses is just the

first step toward utilizing video-EEG data. A potential future direction is to leverage the detected poses for movement assessment, a task analogous to action recognition in the field of computer vision. This involves training a network where the input is a sequence of detected poses, and the output is a description of the movement during that time interval. Such movement analysis can be used for prognostic purposes, particularly in assessing neurological diseases often associated with NICU newborns [43][44].

Method

Dataset Overview

Our NICU newborn pose dataset, WashU-NICU, was collected from neonates admitted to St. Louis Children's Hospital between 2007 and 2011 as part of a prospective video EEG monitoring study [40]. A total of 69 term neonates (≥36 weeks' gestation at delivery) with neonatal encephalopathy were enrolled and underwent continuous vEEG monitoring starting within 48 hours of birth and continuing throughout the duration of therapeutic hypothermia treatment, resulting in over 500 hours of recorded video. Each newborn was critically ill, meeting clinical criteria for moderate to severe hypoxic-ischemic encephalopathy (HIE) based on Eunice Kennedy Shriver National Institute of Child Health and Human Development guidelines [41] or presenting with suspected or confirmed clinical seizures. To construct a diverse and representative dataset, we manually annotated 500 video frames from these recordings, capturing a broad range of poses, lying positions, lighting conditions, and instances of clinician interventions.

The SyRIP dataset [21] is a publicly available collection of infant RGB images, comprising 500 real images and 1200 synthetic images, each annotated with 17 keypoints. For our analysis, we utilized only the real infant images, which were sourced from Google Images and YouTube. These images depict non-hospitalized infants in at-home environments, often dressed in non-standardized clothing and engaged in self-entertainment or resting. The dataset primarily represents infants in casual and uncontrolled settings, which contrasts significantly with the clinical conditions of the NICU environment.

The MINI-RGBD dataset [27] is a synthetic dataset containing both RGB and depth map images of hospitalized infants. Each synthetic image is generated using the Skinned Multi-Infant Linear Model (SMIL) [27], derived from 12 real infants recorded in a children's hospital. Each synthetic infant is depicted in a set of 1000 consecutive frames, primarily in a supine position, with 25 keypoints annotated for each frame. These images are characterized by minimal movement and feature digitally rendered backgrounds that vary between infants but remain consistent across individual sequences. The dataset's synthetic nature and uniform settings limit its representativeness of the more complex NICU environments.

The Baby Pose dataset [28] is a public collection of NICU newborn depth-video recordings, consisting of 16 preterm infants annotated with 12 keypoints. Each neonate is depicted in a sequence of 1800 consecutive depth images, where pixel intensity correlates with object distance from the camera—brighter pixels represent closer

objects, while darker pixels indicate greater distances. The inherent grayscale format of depth images, while suitable for estimating basic body shapes, lacks the detailed textural and contextual information present in conventional RGB images. As a result, models trained solely on depth images may struggle to generalize to RGB formats commonly used in clinical and research settings.

The Mount Sinai NICU, a video-EEG dataset collected at Mount Sinai Hospital in New York City between February 2021 and December 2022 [34]. The dataset includes recordings from 115 critically ill NICU infants, totaling 4705 hours of vEEG data, with 200 images selected for network zero-shot evaluation.

WashU-NICU Dataset Annotation

WashU-NICU dataset originally consisted of unannotated video-EEG files in .avi format with a resolution of 640 x 320. To create an image dataset, we manually captured screenshots from these videos. We built a dataset of 500 images sampled from the videos and split it into 350 images for the training set and 150 images for the validation set. Although 500 images are smaller in scale compared to existing infant pose datasets [21][27][28], we have three key reasons for this approach. First, newborns are often at rest during the long-term monitoring, causing successive video frames to be nearly identical, so sampling a single representative frame during rest intervals is efficient. Second, the lack of keypoint annotations in the videos makes manual labeling of large-scale video frames labor-intensive and time-consuming. Third, our sampled data includes not only normal images of clear and unobstructed newborns but also more challenging

conditions—such as diverse lying positions, various lighting conditions, occlusions, and clinician interventions. Thus, we constructed a representative and effective NICU newborn pose dataset.

We employed LabelMe [45] , an image annotation software, to annotate our dataset following the COCO keypoint format. The annotations include 2D coordinates of 17 keypoints, each associated with visibility indicators, as well as coordinates for bounding boxes and segmentation regions proportional to the image's dimensions. Specifically, each keypoint is annotated with its x and y coordinates relative to the image dimensions. The 17 keypoints include nose, left eye, right eye, left ear, right ear, left shoulder, right shoulder, left elbow, right elbow, left wrist, right wrist, left hip, right hip, left knee, right knee, left ankle, right ankle. In addition, we use a visibility flag (e.g., 0, 1, 2) for every keypoint. A value of 0 denotes that the keypoint is not visible; 1 denotes that the keypoint is visible but not clearly discernible, and 2 denotes the keypoint is clearly visible. One sample image from the WashU-NICU pose dataset is shown in **Figure 4**.



Figure 4 | **Sample image of WashU-NICU pose dataset.** The image on the left depicts the original, unannotated image. The image on the right shows the corresponded labeled image with poses consisting of 17 keypoints (i.e., nose, left eye, right eye, left ear, right ear, left shoulder,

right shoulder, left elbow, right elbow, left wrist, right wrist, left hip, right hip, left knee, right knee, left ankle, right ankle).

NICU Network Analysis

Given an input image $\mathbb{R}^{W\times H\times 3}$, our backbone network extracts the multi-scale feature maps $F_i = \{F_3, F_4, F_5, F_6\}$. Next, we project multi-scale features F_i to a fixed hidden dimension (i.e., 256) and obtain $\{F'_i\}_{i=3}^{5} \in \mathbb{R}^{\frac{H}{2^i}\times\frac{W}{2^i}\times 256}$. Each map is then transformed into a one-dimensional vector by flattening its spatial dimensions, producing $\{F'_i\}_{i=3}^{6} \in \mathbb{R}^{L_i\times 256}$, where $L_i = \frac{H}{2^i}\times\frac{W}{2^i}$. We concatenate these vectors to form $F^* \in \mathbb{R}^{L\times 256}$, where $L = L_3 + L_4 + L_5 + L_6$, which serves as input to the deformable encoder. The deformable encoder containing 6 identical encoder layers iteratively generates the keypoints query $Q_k \in \mathbb{R}^{L\times 256}$. Finally, N pose queries $Q_p \in \mathbb{R}^{256}$ are randomly initialized as input to the deformable decoder. Deformable decoder leverages Q_p and Q_k and generates N pairs of poses $\{P_i\}_{i=1}^{N} \in R^{N\times 2K}$, along with N corresponding confidence scores $\{S_i\}_{i=1}^{N} \in [0,1]^N$. Each pose $P_i = \{(x_i^j, y_i^j)\}_{j=1}^{K}$, represents the 2-D coordinates of the K keypoints in the image and K denotes the total number of detected keypoints. During the trainining and evalution phase, we utilize N pairs of predicted pose to campare with M ground truth newborn poses for loss computation. In the network inference phase, a predefined threshold τ (e.g., $\tau = 0.6$) is used. Predicted poses P_i with the confidence scores S_i exceeding this threshold are selected as the final detected poses.

Backbone Network

We adopt Swin transformer [29] as the backbone network and extracts multi-scale feature maps from its last three stages. We obtain multi-scale feature maps $F_3 \in \mathbb{R}^{\frac{W}{8}\times\frac{H}{8}\times 384}$, $F_4 \in \mathbb{R}^{\frac{W}{16}\times\frac{H}{16}\times 768}$, $F_5 \in \mathbb{R}^{\frac{W}{32}\times\frac{H}{32}\times 1536}$ from a given image.

Deformable Encoder

The deformable encoder consists of N (e.g., $N = 6$) stacked encoder layers. Each encoder layer contains two modules: multi-scale deformable attention (MSDA), and feed forward network (FFN), as shown in **Figure 3c**. Given a multi-scale features vector input F^*, the deformable encoder iteratively updates the features through each encoder layer, ultimately generate keypoint features Q_k. The deformable attention module serves as the primary computational unit in both our deformable encoder and decoder.

Deformable Attention

Multi-scale deformable attention enforces each query element to attend only K (e.g., $K = 4$) sampled keys. Given N query elements, the computational complexity of deformable attention is $O(N \times k)$, where $K \ll N$, resulting in a linear computation complexity. This makes it much more efficient compared to the quadratic complexity (e.g., $O(N\text{^}2)$) of scaled dot-product attention [39]. This low computational complexity further benefits the network by enabling the use of multi-scale feature maps. We provide a visualization of the deformable attention in **Figure 3d**.

We begin by introducing single-scale deformable attention and then extend to multi-scale deformable attention. Given a feature map $F \in \mathbb{R}^{N\times D}$, let f_q be a feature element, where

q denotes the index. We denote $p_q = (p_{qx}, p_{qy})$ to represent 2-D reference point of f_q in F. Then, we denote the equation of the single-scale deformable attention in **Equation 1**:

$$\text{DeformAttn}(f_q, p_q, F) = \sum_{k=1}^{K} A_{qk} \cdot W^V F(p_q + \Delta p_{qk}) \quad (1)$$

A fixed number of sampled keys are selected for attention computation. For example, $K = 4$ denotes that each query element attends to four keys in attention computation, where k is the index of K. The offset Δp_{qk} denotes the 2-d learnable sampling offset w.r.t to the reference point p_q. The relative location of the sampled keys therefore are $p_q + \Delta p_{qk}$. The value of the sampled keys in the feature maps is represented by $F(p_q + \Delta p_{qk})$. In other words, the author in [37] formulate a look-up table retrieval process to locate the sampled keys in the feature maps. Therefore, only Δp_{qk} needs to be updated, saving computation complexity. To obtain $F(p_q + \Delta p_{qk})$, bilinear interpolation is used since $p_q + \Delta p_{qk}$is often a fraction. Furthermore, the attention weight matrix A_{qk} uses linear projection instead of computing dot-product between f_q and the sampled point$\sum_{k=1}^{K} F(p_q + \Delta p_{qk})$ for improved training efficiency.

We then extend the single-scale deformable attention to multi-head, single-scale deformable attention in **Equation 2**:

$$\text{Multihead-DefromAttn}(f_q, p_q, F) = \text{Concat}(head_1, \dots, head_i) \cdot W^O \quad (2)$$

Where $head_i = \text{DeformAttn}(f_q^i, p_q^i, F^i)$, and i index attention head.

Finally, the single-scale deformable attention can be extended to adapt multi-scale feature by simple modification. Given the multi-scale input $F = [F^1, \dots, F^l]$, where $F^l \in \mathbb{R}^{N_i \times D}$, the multi-scale deformable attention is interpreted in **Equation 3**.

$$\text{Multi-Scale DeformAttn}\left(f_q, p_q, F^l\right) = \sum_{l=1}^{L}\sum_{k=1}^{K} A_{lqk} \cdot W^V F^l\left(p_q^l + \Delta p_{lqk}\right) \qquad (3)$$

The **Equation 3** has a new index, l, compared to **Equation 2**, which denotes the scale level. In this case, $p_q^l \in [0,1]^2$ now is a normalized 2-d reference point for each query element at scale l so that p_q can be aligned in each scale level. For example, $(0,0)$ and $(1,1)$ represents top-left and bottom-right location in the feature maps.

Deformable Transformer Decoder

The decoder generates N poses and their confidence score $\{Pose_i\}_{i=1}^N \in \mathbb{R}^{N \times K \times 2}$, $\{Score_i\}_{i=1}^N \in [0,1]^N$, respectively, given N (i.e., $N = 300$) randomly initialized pose query $Q_p \in \mathbb{R}^{N \times 256}$. Each P_i contains K keypoints coordinates $Pose_i = \left\{\left(x_i^j, y_i^j\right)\right\}_{j=1}^K$. The decoder has a similar architecture to the encoder, except it has an additional self-attention and a cross-deformable attention module, as shown in **Figure 3c**. The self-attention allows the embedded pose query Q_p to interact with each other, reasoning a pose-to-pose attention and yielding Q_p'. Then, the cross-deformable attention allows Q_p' to extract multi-scale keypoint features from Q_k to reason full-body pose, resulting in Q_p''. Instead of using $K = 4$ sampling keys as in the deformable encoder, we follow [42] to use $K = 17$ in the deformable decoder to match the number of detected keypoints required by COCO

format. In this way, every query element in pose query Q'_p attends 17 sampled keys in Q_k. Finally, we obtain the predicted poses $\{Pose_i\}_{i=1}^{N}$ and confidence score $\{Score_i\}_{i=1}^{N}$ through a feed forward network (FFN).

Loss Function

During the training phase, we follow to use the Hungarian algorithm [37] that determines the unique assignments between the sets of N predicted and n ground-truth poses. We follow [16] to jointly employ L_1loss and Object Keypoint Similarity (OKS) loss for pose regression. The OKS loss is defined in **Equation 4**:

$$L_{oks}(Pose_i, Pose_{gt}) = \sum_{j=1}^{K} \frac{\exp(\| Pose_i - Pose_{gt} \|/2s^2k_i^2)\, \delta(v_i > 0)}{\sum_{j=1}^{K} \delta(v_i > 0)} \qquad (4)$$

where $\| Pose_i - Pose_{gt} \|$ is the Euclidian distance between the predicted pose and the ground-truth pose, v_i is the visibility of the ground-truth keypoint, s is the object scale, and k_i is the per-keypoint constant that controls fall-off. Besides, we also use Focal loss (L_{cls}) to for the class classfication.

Thereby, we define the overall loss function in **Equation 5**:

$$L = \lambda_1 L_{oks} + \lambda_2 L_1 + L_{cls} \qquad (5)$$

where λ_i are the loss weights.

Reference

[1]. Pouppirt NR, Martin V, Pagnotto-Hammitt L, Spittle AJ, Flibotte J, DeMauro SB. The General Movements Assessment in Neonates With Hypoxic Ischemic Encephalopathy. Journal of Child Neurology. 2021;36(8):601-609. doi:10.1177/0883073820981515

[2]. Einspieler C, Bos AF, Libertus ME, Marschik PB. The General Movement Assessment Helps Us to Identify Preterm Infants at Risk for Cognitive Dysfunction. Front Psychol.

2016 Mar 22;7:406. doi: 10.3389/fpsyg.2016.00406. PMID: 27047429; PMCID: PMC4801883.
[3]. Lester, Barry M., et al. "Assessment and evaluation of the high risk neonate: the NICU Network Neurobehavioral Scale." JoVE (Journal of Visualized Experiments) 90 (2014): e3368.
[4]. Seesahai J, Luther M, Church PT, Maddalena P, Asztalos E, Rotter T, Banihani R. The assessment of general movements in term and late-preterm infants diagnosed with neonatal encephalopathy, as a predictive tool of cerebral palsy by 2 years of age-a scoping review. Syst Rev. 2021 Aug 12;10(1):226. doi: 10.1186/s13643-021-01765-8. PMID: 34384482; PMCID: PMC8359053.
[5]. Barnes F, Graham L, Loganathan P, Nair V. General Movement Assessment Predicts Neuro-Developmental Outcome in Very Low Birth Weight Infants at Two Years - A Five-Year Observational Study. Indian J Pediatr. 2021 Jan;88(1):28-33. doi: 10.1007/s12098-020-03365-1. Epub 2020 Jun 3. PMID: 32488806.
[6]. Ferrari F, Cioni G, Prechtl HF. Qualitative changes of general movements in preterm infants with brain lesions. Early Hum Dev. 1990 Sep;23(3):193-231. doi: 10.1016/0378-3782(90)90013-9. PMID: 2253580.
[7]. Sandoval Karamian AG, Wusthoff CJ. Current and Future Uses of Continuous EEG in the NICU. Front Pediatr. 2021 Nov 3;9:768670. doi: 10.3389/fped.2021.768670. PMID: 34805053; PMCID: PMC8595393.
[8]. Pressler RM, Cilio MR, Mizrahi EM, Moshé SL, Nunes ML, Plouin P, Vanhatalo S, Yozawitz E, de Vries LS, Puthenveettil Vinayan K, Triki CC, Wilmshurst JM, Yamamoto H, Zuberi SM. The ILAE classification of seizures and the epilepsies: Modification for seizures in the neonate. *Position paper by the ILAE Task Force on Neonatal Seizures. Epilepsia.* 2021 Mar;62(3):615-628. doi: 10.1111/epi.16815. Epub 2021 Feb 1. PMID: 33522601.
[9]. Temko, Andriy, et al. "EEG-based neonatal seizure detection with support vector machines." Clinical Neurophysiology 122.3 (2011): 464-473.
[10]. O'Shea, Alison, et al. "Neonatal seizure detection from raw multi-channel EEG using a fully convolutional architecture." Neural Networks 123 (2020): 12-25.
[11]. Siddiqui, Mohammad Khubeb, et al. "A review of epileptic seizure detection using machine learning classifiers." Brain informatics 7.1 (2020): 5.
[12]. Abbasi, H., Mollet, S.R., Williams, S.A. et al. Deep-learning for automated markerless tracking of infants general movements. Int. j. inf. tecnol. 15, 4073–4083 (2023). https://doi.org/10.1007/s41870-023-01497-z
[13]. Luca Lonini, Yaejin Moon, Kyle Embry, R. James Cotton, Kelly McKenzie, Sophia Jenz, Arun Jayaraman; Video-Based Pose Estimation for Gait Analysis in Stroke Survivors during Clinical Assessments: A Proof-of-Concept Study. Digit Biomark 10 May 2022; 6 (1): 9–18.
[14]. Jaroensri, R., Zhao, A., Balakrishnan, G., Lo, D., Schmahmann, J.D., Durand, F. & Guttag, J.. (2017). A Video-Based Method for Automatically Rating Ataxia. Proceedings of the 2nd Machine Learning for Healthcare Conference, in Proceedings of Machine Learning Research 68:204-216 Available from https://proceedings.mlr.press/v68/jaroensri17a.html.
[15]. Jaroensri, Ronnachai, et al. "A video-based method for automatically rating ataxia." Machine Learning for Healthcare Conference. PMLR, 2017.
[16]. Shi, Dahu, et al. "End-to-end multi-person pose estimation with transformers." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022.

[17]. Sun, Ke, et al. "Deep high-resolution representation learning for human pose estimation." Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2019.
[18]. Cheng, Bowen, et al. "Higherhrnet: Scale-aware representation learning for bottom-up human pose estimation." Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2020.
[19]. Yuan, Yuhui, et al. "Hrformer: High-resolution transformer for dense prediction." arXiv preprint arXiv:2110.09408 (2021).
[20]. Geng, Zigang, et al. "Bottom-up human pose estimation via disentangled keypoint regression." Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2021.
[21]. Huang, Xiaofei, et al. "Invariant representation learning for infant pose estimation with small data." *2021 16th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2021)*. IEEE, 2021.
[22]. Cao, Xu, et al. "AggPose: Deep aggregation vision transformer for infant pose estimation." arXiv preprint arXiv:2205.05277 (2022).
[23]. Groos, Daniel, et al. "Towards human performance on automatic motion tracking of infant spontaneous movements." arXiv preprint arXiv:2010.05949 (2020).
[24]. Moccia, Sara, et al. "Preterm infants' pose estimation with spatio-temporal features." IEEE Transactions on Biomedical Engineering 67.8 (2019): 2370-2380.
[25]. Q. Wu et al., "Supine Infant Pose Estimation via Single Depth Image," in IEEE Transactions on Instrumentation and Measurement, vol. 71, pp. 1-11, 2022, Art no. 5013711, doi: 10.1109/TIM.2022.3178693.
[26]. Malfilâtre G, Mony L, Hasaerts D, Vignolo-Diard P, Lamblin MD, Bourel-Ponchel E. Technical recommendations and interpretation guidelines for electroencephalography for premature and full-term newborns. Neurophysiol Clin. 2021 Jan;51(1):35-60. doi: 10.1016/j.neucli.2020.10.005. Epub 2020 Nov 7. PMID: 33168466.
[27]. Hesse, Nikolas, et al. "Computer vision for medical infant motion analysis: State of the art and rgb-d data set." *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*. 2018.
[28]. Migliorelli L, Moccia S, Pietrini R, Carnielli VP, Frontoni E. The babyPose dataset. *Data Brief*. 2020 Sep
[29]. Caron, Mathilde, et al. "Emerging properties in self-supervised vision transformers." *Proceedings of the IEEE/CVF international conference on computer vision*. 2021.
[30]. Sun, Ke, et al. "Deep high-resolution representation learning for human pose estimation." *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2019.
[31]. Yuan, Yuhui, et al. "Hrformer: High-resolution transformer for dense prediction." *arXiv preprint arXiv:2110.09408* (2021).
[32]. Cheng, Bowen, et al. "Higherhrnet: Scale-aware representation learning for bottom-up human pose estimation." *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2020.
[33]. Geng, Zigang, et al. "Bottom-up human pose estimation via disentangled keypoint regression." *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2021.
[34]. Gleason A, Richter F, Beller N, Arivazhagan N, Feng R, Holmes E, Glicksberg BS, Morton SU, Vega-Talbott M, Fields M, Guttmann K, Nadkarni GN, Richter F. Accurate prediction of neurologic changes in critically ill infants using pose AI. medRxiv [Preprint]. 2024 Apr 19:2024.04.17.24305953. doi: 10.1101/2024.04.17.24305953. PMID: 38699362; PMCID: PMC11064996.

[35]. Dosovitskiy, Alexey, et al. "An image is worth 16x16 words: Transformers for image recognition at scale." arXiv preprint arXiv:2010.11929 (2020).
[36]. McInnes, Leland, John Healy, and James Melville. "Umap: Uniform manifold approximation and projection for dimension reduction." *arXiv preprint arXiv:1802.03426* (2018).
[37]. Zhu, Xizhou, et al. "Deformable detr: Deformable transformers for end-to-end object detection." *arXiv preprint arXiv:2010.04159* (2020).
[38]. Xia, Zhuofan, et al. "Vision transformer with deformable attention." Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2022.
[39]. Vaswani, Ashish, et al. "Attention is all you need." Advances in neural information processing systems 30 (2017).
[40]. Srinivasakumar P, Zempel J, Trivedi S, Wallendorf M, Rao R, Smith B, Inder T, Mathur AM. Treating EEG Seizures in Hypoxic Ischemic Encephalopathy: A Randomized Controlled Trial. Pediatrics. 2015 Nov;136(5):e1302-9. doi: 10.1542/peds.2014-3777. Epub 2015 Oct 19. PMID: 26482675.
[41]. Shankaran, Seetha, et al. "Whole-body hypothermia for neonates with hypoxic–ischemic encephalopathy." New England Journal of Medicine 353.15 (2005): 1574-1584.
[42]. Lin, Tsung-Yi, et al. "Microsoft coco: Common objects in context." *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*. Springer International Publishing, 2014.
[43]. Leo, M.; Bernava, G.M.; Carcagnì, P.; Distante, C. Video-Based Automatic Baby Motion Analysis for Early Neurological Disorder Diagnosis: State of the Art and Future Directions. *Sensors* 2022, *22*, 866. https://doi.org/10.3390/s22030866
[44]. Sandoval Karamian AG, Wusthoff CJ. Current and Future Uses of Continuous EEG in the NICU. Front Pediatr. 2021 Nov 3;9:768670. doi: 10.3389/fped.2021.768670. PMID: 34805053; PMCID: PMC8595393.
[45]. Wada, K. Labelme: Image Polygonal Annotation with Python [Computer software]. https://doi.org/10.5281/zenodo.5711226

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupplementaryMaterials.docx