Abstract—Infant pose monitoring during sleep has multiple applications in both healthcare and home settings. In a healthcare setting, pose detection can be used for region-of-interest (ROI) detection and movement detection for noncontact-based monitoring systems. In a home setting, pose detection can be used to detect sleep positions, which has shown to have a strong influence on multiple health factors. However, pose monitoring during sleep is challenging due to heavy occlusions from blanket coverings and low lighting. To address this, we present a novel dataset, Simultaneously-collected multimodal Mannequin Lying pose (SMaL) dataset, for under-the-cover infant pose estimation. We collect depth and pressure imagery of an infant mannequin in different poses under various cover conditions. We successfully infer full body pose under the cover by training state-of-the-art pose estimation methods and leveraging existing multimodal adult pose datasets for transfer learning. We demonstrate a hierarchical pretraining strategy for transformer-based models to significantly improve performance on our dataset. Our best-performing model was able to detect joints under the cover within 25 mm 86% of the time with an overall mean error of 16.9 mm. Data, code, and models are publicly available at https://github.com/DanielKyr/SMaL.

Index Terms—Depth sensing, infant pose estimation, multimodal data, pressure mapping, pretraining, vision transformers (ViTs).

I. INTRODUCTION

In the neonatal intensive care unit (NICU), continuous monitoring is required due to the neonate’s fragile health condition. This often involves the use of wired sensors placed on the chest area and limbs to measure and track a newborn’s physiologic signals. However, false alarms can be generated due to motion artifacts created at the skin–electrode interface during patient movement [1]. Studies have shown that 87.5% of alarms in the NICU are false alarms, with a majority caused by movements of the infants [2]. In addition, the adhesives used for the electrodes can irritate the fragile neonate’s skin [3]. This study is part of a larger research initiative that aims to address these issues by developing novel nonobtrusive neonatal patient monitoring systems based on pressure-sensitive mats (PSMs) and color/depth (RGB-D) video for the purpose of monitoring vital signs and detecting patient movements for the reduction of false alarms.

Initial studies showed that it is possible to detect patient movement to gate false alarms arising from motion artifacts [4]. These preliminary results showed that general motion detection was insufficient to create a robust system and that more precise movement tracking was required. Precisely detecting the neonate’s individual limb motion through pose monitoring would allow a system to better associate false alarms caused on specific sensors with the movement of the limb with which the sensor was placed. Pose monitoring can also be beneficial to noncontact vital sign monitoring techniques, which rely on region-of-interest (ROI) detection [5]. One can extract relevant anatomical regions from pose, such as the chest, which then can be used for respiration rate (RR) estimation using color, depth, or pressure-based technologies [6], [7]. In addition, since these methods are sensitive to motion artifacts [5], pose detection could also be used to reject estimates where the patient is moving significantly. Outside of a healthcare setting, monitoring an infant’s pose during sleep can provide critical information about an infant’s sleeping position. The sleeping position of an infant has a strong influence on multiple health factors, such as sudden infant death syndrome (SIDS) [8], cerebral oxygenation [9], heart rate variability (HRV) [10], and obstructive sleep apnea [11].

It is clear that pose monitoring is a critical part of NICU and home-monitoring systems. However, standard pose estimation techniques are not suited for monitoring during sleep since the environment introduces adverse vision conditions: heavy occlusions due to blanket coverings and low lighting (including complete darkness for premature infants in isolettes). These challenging conditions can be mitigated by employing additional sensing modalities other than color imagery, such as contact pressure and depth data. To partially address this challenge, Liu et al. [12] released an in-bed human pose dataset, called Simultaneously-collected multimodal Lying Pose (SLP). The SLP dataset contains adult pose data collected under three conditions: no cover, a thin sheet, and a thick blanket, using four sensing modalities: color imaging, thermal imaging, depth sensing, and pressure sensing. However, this dataset only contains adult subjects and does not include...
infant subjects. Most existing human pose estimation methods are targeted toward adults whose anatomical proportions and distribution of body poses differ from infants due to muscle tone and joint flexibility. In addition, there are differences in contact pressure patterns between adults and infants on the PSM-based on their anatomical weight distribution. Due to these differences, it has been shown that when adult pose estimation methods are employed on infant images, the localization performance drops significantly [13]. Therefore, there is a need for an equivalent infant-specific dataset to develop pose estimation models that are accurate for infants. Collecting an equivalent dataset using real infants would not be feasible since infants will not hold their position between cover conditions; instead, a poseable infant mannequin can be used. With this in mind, we introduce the Simultaneously-collected multimodal Mannequin Lying pose (SMaL) dataset. This dataset contains a set of 300 unique poses under three cover conditions using three sensor modalities: color imaging, depth sensing, and pressure sensing. Since we are leveraging a mannequin, thermal imaging is not relevant. The SMaL dataset represents the first multimodal dataset for infant pose estimation and the first dataset to explore under-the-cover pose estimation for infants. The exploration of multimodal data is important as fusing pressure and depth imagery allows for each modality to compensate for the weaknesses of the other. For example, pressure images provide little data relating to free-moving limbs that produce negligible contact pressure. Conversely, depth information is limited when the patient is heavily occluded.

The SMaL dataset is a relatively small dataset when compared to the SLP dataset. This represents a common challenge in the infant pose estimation field—when compared to adult datasets, infant data is limited. A common approach to address this small data problem is the use of transfer learning. The two common sources of additional data are adult datasets and synthetic datasets [14]. To improve performance on our SMaL dataset, we leverage two additional data sources for transfer learning. The first is the SLP dataset, which contains labeled adult depth and PSM images [12], which we use for additional training data for pose estimation. The second is BodyPressureSD that contains simulated depth and pressure images of adults in lying positions under the blanket covering [15]. With the recent surge in vision transformer (ViT) models and their superior performance, we utilize these data as a source for a self-supervised pretraining task for a ViT pose estimation model. We employ the methods of Reed et al. [16] for hierarchical pretraining and propose our own hierarchical pretraining strategy for infant pose estimation. To the best of our knowledge, this work represents the first study that employs a hierarchical pretraining strategy for pose estimation.

This work extends a previous work titled “Transfer Learning Approaches for Neonate Head Localization from Pressure Images” [17], which demonstrated: 1) successful transfer of annotations across domains (i.e., RGB to PSM) and 2) improved performance on the head detection task when using a convolutional neural network (CNN) backbone previously trained on adult pressure imagery. The present study extends these approaches to estimate full body pose, with 14 joints, including the head, from a combination of depth and pressure imagery. We successfully trained pose estimation models to infer under the cover infant pose from the depth and pressure imagery. We also demonstrate a masked autoencoding (MAE) hierarchical pretraining strategy for ViT that significantly improves accuracy on the SMaL dataset. The main contributions of the current works are summarized as follows.

1) Present a mannequin infant in-bed pose dataset, SMaL, with simultaneously-collected multimodal imagery: RGB, depth, and pressure under different cover conditions.

2) Train and compare various state-of-the-art pose estimation models, including both CNN- and transformer-based models, on the SMaL dataset, while leveraging adult data for transfer learning.

3) Establish, for the first time, the benefit of a hierarchical pretraining strategy for pose estimation using MAE for ViT-based models.

Section II covers related literature. Section III presents the data collection methodology for the SMaL dataset. Section IV describes the ViT architecture and the hierarchical pretraining paradigm. Section V presents the pose estimation model architectures and the training strategies used for each, as well as how we evaluate the methods. This is followed by the results and discussion in Section VI.

II. RELATED WORKS

A. In-Bed Adult Pose Estimation

Although there is a multitude of works addressing human pose estimation, only a few address the resting condition where the human is lying in a bed. This is first due to the fact that, until recently, there was a lack of available large-scale datasets. Second, although RGB data has been leveraged for at-rest posture estimation, these methods are constrained to well-illuminated environments and cases with little-to-no occlusion [18]. To overcome these challenges, researchers have leveraged additional modalities for this task. Pressure data has been used extensively for the estimation of 2-D pose [19], 3-D pose [20], [21], and even complete 3-D human pose and body shape [22]. However, when solely using pressure images for pose estimation, ambiguities arise from free-moving limbs producing negligible contact pressure [20], [23]. Depth data has also been leveraged for estimating human pose during sleep. Clever et al. [15] were able to estimate full 3-D adult pose and body shape and estimate a corresponding pressure image under the cover using depth imagery. To train their model, they generated synthetic data via a soft-body physics simulation of a human body, a mattress, a pressure-sensing mat, and a covering blanket. Their publicly available dataset, BodyPressureSD, contains approximately 100 k simulated depth and pressure images of adults in lying positions under the blanket covering. Currently, the SLP dataset is the benchmark dataset for in-bed pose estimation [12]. It is a large-scale dataset with over 100 subjects and nearly 15 000 pose images, which is comparable to other well-known general-purpose human pose datasets [12]. The posture coverage evenly comprises supine, left-, and right-side sleep posture categories. Each
subject was asked to perform 15 poses for each category under three cover conditions using four imaging modalities simultaneously. Given the multimodal nature of their dataset, they demonstrated that the pressure data was complementary to other modalities; by stacking pressure data with other modalities, such as thermal, depth, or both, a performance improvement was achieved over their single modality counterparts. The size of the dataset and its comprehensive set of modalities makes the SLP dataset versatile for different vision tasks, exemplified by the recent works that leverage this dataset [24], [25], [26], [27].

B. Infant Pose Estimation

There is a significant amount of research focused on infant pose estimation. The main motivation behind developing these methods is to create automatic systems that can perform general movements assessment (GMA) for the diagnosis of cerebral palsy and other developmental disabilities. RGB and depth are the two main modalities utilized for this task. There are various works that investigate RGB images for infant pose detection describing a wide range of infant pose estimation methods, including both CNN-based [13] and transformer-based [28] methods. Since there are not many public infant datasets when compared to adult datasets, some works opt to use synthetic data for training [14], [29].

Rather than RGB images, several studies have opted for depth imagery, due to its privacy-preserving nature. Moccia et al. [30] released a dataset of 16 k annotated depth images from the 27 patients in the NICU for infant pose estimation, called babyPose. They also trained a CNN using spatiotemporal features on consecutive depth frames. Wu et al. [33] curated a dataset of 50 k annotated depth images from the 27 patients in the NICU. They labeled this data semi-automatically by running existing pose estimators on the RGB images and transferring the annotations to the depth stream. To our knowledge, no work has attempted infant pose estimation from contact pressure imagery; however, some works have explored posture recognition [31], [32]. Our previous work, explored head localization from pressure imagery but not the full pose [17].

Although, as mentioned, there are depth-based infant pose estimation datasets [30], [33], these have focused on applications for GMA and do not consider scenarios in which the infant is fully occluded with coverings. In addition, no pressure-based or multimodal data exists for infant pose estimation. In this work, we aim to fill these gaps through our mannequin infant dataset, SMaL, which includes simultaneously-collected RGB, depth, and pressure under different cover conditions. This is the first dataset to explore lying infant pose using multiple modalities and under heavy occlusions.

III. SMaL Dataset

A. Experimental Setup

The data collection setup includes an Intel RealSense D435 camera to capture color and depth video data (B) (Intel Corporation, Santa Clara, CA, USA) and a LX100:100.100.05 PSM (XSensor Technology Corporation, Calgary, AB, Canada). The D435 camera has a resolution of up to 1280 × 720 px for RGB and 1920 × 1080 px for depth. It uses stereo vision to calculate the depth and an infrared (IR) projector to project a non-visible static IR pattern to improve depth accuracy in scenes with low texture. The depth accuracy is ±2% at 2 m. The PSM sensor has a spatial resolution of 5.08 mm with an overall sensing area of 50.8 × 50.8 cm², made up of a grid of 100 × 100 sensels. It captures ballistographic signals ranging from 0.0 to 3.41 N/cm². The camera was placed overhead of the PSM at a distance of 0.5 m with a small foam mattress underneath. The mannequin used was the StandInBabyB. mannequin (StandInBaby, Springfield Lakes, QLD, Australia). It features articulated joints that cover a full range of natural motion, accurate size, and weight distribution representing the 50th percentile of a newborn infant (50-cm length, 6.8-lbs weight). The photograph of the experimental setup is shown in Fig. 1, with sample data from the depth and PSM streams.

B. Multimodal Registration

A grid of visible landmarks was drawn on the PSM to obtain reliable registration between the RGB and PSM streams, as demonstrated in [23]. The landmarks form a 6-by-6 grid with a 9-cm separation between each point. The true locations of these points are known in the PSM plane by their position relative to the border marked by the manufacturer. Since these points are visible by the RGB-D video camera, their location in the video plane is precisely known. Before any data collection session, the location of the visible landmarks was identified in the RGB stream. With the known location of these landmarks in the PSM stream, homography was calculated using a random sample consensus (RANSAC)-based robust method. The depth stream was aligned to the PSM by first converting the RGB coordinates of the landmarks to depth.
coordinates using the camera parameters and then calculating a new homography matrix for the depth stream.

C. Data Collection

The infant mannequin pose data were collected using custom software with a graphical user interface (GUI). The software live-streamed the RGB and depth data and was used to annotate registration points as well as pose annotation during data collection. The PSM data were recorded using the XSensor Pro 8 software and were time-aligned to the camera data following data collection. For each pose, the mannequin was placed in a novel position, the pose was annotated on the RGB stream and an image was captured for each of three cover conditions: uncovered, thin cover, and thick cover. The thin cover was a bed sheet that was approximately 1 mm in thickness, while the thick cover was a blanket that was approximately 5 mm in thickness. The pose annotation followed that of the SLP dataset, focusing on the major limbs for a total of 14 joints: Ankle-R, Knee-R, Hip-R, Hip-L, Knee-L, Ankle-L, Wrist-R, Elbow-R, Shoulder-R, Shoulder-L, Elbow-L, Wrist-L, Thorax, and Head Top, where L and R stand for the left and right side, respectively. Poses were categorized into three categories: supine position, left side, and right side, to match the SLP dataset. For each category, 100 poses were captured for a total of 300 unique poses each under the three cover conditions. To transfer the annotations to the PSM from RGB, the homography matrix calculated from the RGB-PSM registration was used. This annotation was transferred to all the cover conditions. The depth image was then aligned using the depth to PSM homography. The final data contained the depth-aligned image, the pressure image, and the pose annotation in PSM coordinates. The final images were sized to $224 \times 224$ to be compatible with the ViT backbones. Fig. 2 displays some examples from the dataset.

IV. Vision Transformers

A. Pretraining and Finetuning

Pretraining refers to training a model on a pretext task using a large quantity of data, often in the millions of samples, to learn general knowledge stored in the form of learned parameters. Finetuning refers to transferring this general knowledge to application-specific downstream tasks. This knowledge transfer is accomplished by initializing model parameters with the parameters learned through pretraining and then adjusting them via supervised learning on annotated data. Pretraining can be accomplished through supervised learning, which trains models using labels gathered through annotation; pretraining can also be accomplished through self-supervised learning (SSL), which trains models using labels generated through an SSL algorithm. For example, labels can be generated by hiding portions of input data from the model and then training the model to reconstruct the hidden content. This is a popular class of SSL algorithms often referred to as reconstructive learning and is recently exemplified by MAE [34], which we leverage in Section V. A model architecture that fits incredibly well in the “pretraining and finetuning” paradigm is the transformer.

B. Transformer Architecture

Introduced by Vaswani et al. [35], the transformer is an attention-based model architecture used across many applications. When a “vanilla” transformer is used in computer vision applications, it is called a ViT [36]. After adapting
pretrained ViTs to specific tasks, ViTs achieve state-of-the-art performance in image classification [34], [37], semantic segmentation [37], object detection [38], and, importantly for our work, pose estimation [39].

ViTs first process an image into a set of “patches”; then, these patches are processed by a transformer model, which outputs patch encodings. Processing an image into a set of patches is accomplished by splitting the image into small squares, usually 16 by 16 pixels, resulting in patch tensors of shape $16 \times 16 \times 3$. Since our images have a height and width of 224 pixels, dividing by 16 pixels per patch leaves us with a 14 by 14 grid of patches, which are then arranged in a set of 196 total patches. Next, these patch tensors are flattened, resulting in 768-d feature vectors, still retaining the original pixel values. In the penultimate patchification step, each of the 768-d feature vectors is linearly projected to the transformer model width. In this work, we use ViT-Base models, which have a model width of 768. Thus, this linear projection outputs 768-d feature vectors that we call patch representations. In the final patchification step, each patch representation gets added to a unique sinusoidal position embedding that represents the location of each patch within the image. Now, each patch representation contains both content and positional information.

Next, this set of patch representations is processed by a transformer model composed of stacked transformer blocks. Each transformer block is composed of a self-attention sublayer and a feedforward sublayer. Self-attention sublayers learn to exchange features between patches, and feedforward sublayers learn to transform the features of individual patches. The ViT-Base architecture employs 12 transformer blocks to encode each patch with contextual “higher-level” features. The term “ViT encoder” refers to both patchification and transformer modules. The network architecture for a ViT backbone is shown in the middle section of Fig. 3.

C. Masked Autoencoding

MAE is an SSL algorithm that pretrains ViTs by encoding a small portion of patches (referred to as visible patches) and then employs a ViT decoder to reconstruct the hidden patches. In more detail, MAE first encodes a randomly sampled 25% of image patches using a ViT encoder and holds the remaining 75% of patches aside. The ViT decoder receives visible patch encodings and mask embeddings. These embeddings represent the hidden patches and can be thought of as placeholders for the decoder to write. Before the decoder processes these inputs, another set of sinusoidal position embeddings are added to both visible patch encodings and mask embeddings; these position embeddings represent the locations of patches within the original image. Given these inputs, the decoder outputs an image prediction, matching the image’s shape of $224 \times 224 \times 3$. The loss is then computed by the mean squared error between the original held-out patches and the decoder’s predictions of the held-out patches. The decoder’s predictions of visible patches are ignored. Both encoder and decoder are trained end-to-end; after pretraining, the decoder is typically discarded, and only the ViT encoder is leveraged through finetuning on downstream tasks. When pretraining with MAE, a smaller decoder size is typically chosen to reduce the computational costs of pretraining; specifically, they chose a decoder model width of 512 and two transformer blocks. This asymmetric encoder–decoder design reduces pretraining FLOPs by approximately $3 \times$ [34]. The baseline ViT encoder we use in this article was pretrained by He et al. [34] for 1600 epochs on the ImageNet-1k dataset.

D. Hierarchical Pretraining

The similarity between upstream (pretraining) and downstream (finetuning) tasks is a leading indicator of downstream performance [40]. In our case, we do not have access to a model pretrained on fused depth and pressure data. Consequently, any pretrained model, we initialize our model with, will result in dissimilar upstream and downstream tasks; this is not ideal for knowledge transfer. Facing the same challenge, Reed et al. [16] propose hierarchical pretraining, which progressively pretrains models on data closer to their downstream
data, starting from a model pretrained on ImageNet. We leverage hierarchical pretraining by continuing to pretrain our ViT encoder on in-domain (i.e., fused depth and pressure) data using the MAE SSL algorithm. This results in three pretrained models that have been specialized to our domain by pretraining on: 1) in-domain simulated data; 2) in-domain real data; and 3) first in-domain simulated, then in-domain real data. We compare these three models to our baseline model, which did not undergo further in-domain pretraining. For our first two specialized models, we pretrain on in-domain—either simulated or real—data for 150 epochs, initializing our ViT encoder from MAE’s pretrained model parameters and our ViT decoder randomly. Our decoder must be initialized randomly because the MAE authors do not share their decoder model parameters. For our third specialized model, we first pretrain on in-domain simulated data for 150 epochs, initializing our ViT encoder from MAE’s pretrained model parameters and our ViT decoder randomly. Then, we continue to pretrain on in-domain real data for 150 epochs, initializing our ViT encoder and decoder model parameters from the prior stage. This leaves us with three specialized ViT encoders, which we leverage through finetuning on annotated pose estimation data.

V. UNDER THE COVER INFANT POSE ESTIMATION

We train and test state-of-the-art pose estimation methods on our SMaL dataset. We compare various model architectures and various approaches to hierarchical pretraining for the pure ViT models.

A. Datasets

1) BodyPressureSD: BodyPressureSD contains 97,495 unique body shapes, poses, and image samples. Each sample consists of a human model with a realistic lying pose. From this model, simulated depth images in covered and uncovered settings and a corresponding physics-based pressure image were simulated. The depth image shape is $128 \times 54$ and the pressure image shape is $64 \times 27$. They are spatially aligned and can be stacked. Both uncovered and covered conditions were used, for a total dataset size of 194,990 images. A sample set of images from BodyPressureSD are shown in Fig. 4.

2) SLP: SLP contains 4590 real adult in-bed poses. For each sample, there were three cover conditions (i.e., uncovered, thin cover, and thick cover). The depth image shape is $512 \times 424$ and the pressure image shape is $192 \times 84$. The images were aligned using the provided homography. The dataset was split into a training set of 4050 and the remaining was used as a validation set. All cover conditions were used for training for a total of 12,150 images; only covered conditions were used for validation and testing for a total of 120 images for each.

3) SMaL: SMaL contains 300 mannequin poses. For each sample, there were three cover conditions (i.e., uncovered, thin cover, and thick cover). The depth image size is $640 \times 480$ and the pressure image shape is $100 \times 100$. The images were aligned using the registration process described in Section III-B. The dataset was split into five folds of 60 each for cross-validation, due to the small size of the dataset. Three folds were used for training, one as validation for early stopping, and one for testing. All cover conditions were used for training for a total of 540 images; only covered conditions were used for validation and testing for a total of 120 images for each.

B. Training and Evaluation

Below are the transfer learning steps and their corresponding data distributions.

1) ViT backbone (ImageNet-1k distribution via MAE).
2) Specialized ViT backbone (BodyPressureSD distribution via MAE).
3) Specialized ViTPose (SLP distribution via finetuning).
4) Final ViTPose model (SMaL distribution via finetuning).

Typically, models are directly fine-tuned from a model pretrained on ImageNet (i.e., from step 1 directly to step 4). We introduce steps 2 and 3 that leverage self-supervised MAE training and finetuning, respectively, using additional adult pose datasets. The synthetic BodyPressureSD dataset was used solely for pretraining the ViT backbone. Based on experiments from Reed et al. [16], this dataset surpasses the minimum size requirement for use as a secondary pretraining dataset. In addition, exact 2-D pose coordinates are not available for this dataset, excluding it from use during the finetuning stages. The SLP dataset was used for both pretraining the ViT backbone and finetuning.
TABLE I
TRAINING PARAMETERS

| Parameters      | SLP | SMAIL |
|-----------------|-----|-------|
| Batch Size      | 256 | 16    |
| Learning Rate   | 1e-3| 1e-4  |
| Training Iterations | 50  | 50   |
| Warm Up Iterations | 5   | 5    |
| Augmentations   | SLP | SMAIL |
| Rotation Factor | <30°| <2°   |
| Scaling Factor  | <0.25| 0    |
| Do Occlusion    | 50% | 0     |
| Color Scaling   | <0.2| 0     |
| Do H-Flip       | 50% | 50%   |

1) Training Settings: Finetuning the models was done from scratch and comprised of steps 3 and 4. Step 3 was finetuning the training set of SLP and using the validation for early stopping. Step 4 was continuing the finetuning using the SMAIL training set with early stopping on the validation set and a final evaluation on the test set. The second stage of finetuning was repeated for the fivefold cross-validation. The input image for all the models had a shape with $224 \times 224$ with two channels: one for depth and one for pressure. Since the modalities are spatially aligned, channel stacking is possible. The depth channel was duplicated to be compatible with pretrained ImageNet backbones. The joint locations were transformed into heatmaps by setting the joint location to the maximum point of a 2-D Gaussian distribution, with a sigma equal to 1. The heatmap method has become a standard for human pose estimation and is employed by most of the state-of-the-art methods. These heatmaps have a shape of $56 \times 56$ to match the downsampling factor of four that is commonly used. Therefore, the final targets had a shape of $56 \times 56 \times 14$, one channel for each joint. The loss function used was mean squared error, with the AdamW optimizer, five warmup epochs, and half-cycle cosine learning rate decay. Major augmentations were applied when training SLP to prevent overfitting. Only slight rotation and horizontal flipping were used during SMAIL training as these are the only variations that exist in the dataset given the limited dataset size. A summary of the training settings for the SLP and SMAIL datasets are shown in Table I.

2) Evaluation: Two metrics were used to evaluate the models on the SMAIL dataset: percent of correct keypoints (PCK) and the normalized mean error (NME). To count as a correct keypoint for PCK, the predicted keypoint had to be within 0.05 of the normalized distance from the ground-truth keypoint. A normalized distance threshold can be used because all images are at the same scale after applying the camera to PSM homography. Since the real scale of the PSM is known, a 0.05 threshold is equivalent to 25.4 mm (each sensel corresponds to 5.08 mm). Equivalently, the NME is multiplied by the real-world scale to measure the true error in mm. Only the PCK is used for early stopping on the validation set during both stages of finetuning. Both metrics are used when evaluating the test set.

C. Model Architectures

Three models were tested to cover a diverse set of architectures: 1) HRNET which uses only CNNs; 2) transpose which uses a CNN-transformer hybrid; and 3) ViTPose which uses a plain vanilla ViT.

1) HRNet: HighResolution Net (HRNet) is a purely CNN-based approach to pose estimation. HRNet is able to maintain high-resolution representations throughout the whole neural network, as opposed to previous methods which recover high-resolution representations from low-resolution representations. This is possible through the parallel nature of HRNet which exchanges the information across parallel multiresolution subnetworks over and over and finally recovers joint estimates from the high-resolution network. There are two HRNet variants: one small net and one big net: HRNet-W32 and HRNet-W48, where 32 and 48 represent the widths of the high-resolution subnetworks in the last three stages, respectively [41].

2) TransPose: The TransPose model consists of three components: a CNN backbone to extract low-level image features, a transformer model to capture long-range spatial interactions between feature vectors across locations, and a head to predict the keypoint heatmaps. Only the initial several layers of the original ImageNet pretrained CNNs are used to extract features from images, the parameters numbers of which are only about 5.5% and 25%, for ResNet and HRNet, respectively. After passing the image through a CNN, the feature maps are flattened and used as input to the transformer which arranges its output into the 2-D-structure heatmaps. The variants of transpose include different backbones (i.e., TransPose-R has a ResNet and TransPose-H has HRNet) and a different number of transformer blocks (i.e., TransPose-H-A4 has four blocks and TransPose-H-A6 has 6). TransPose outperformed both HRNet variants on the common objects in context (COCO) dataset [42].

3) ViTPose: ViTPose employs a pretrained ViT encoder and a lightweight decoder for pose estimation. The ViT encoder produces a 768-d representation of each patch. After arranging the encoder’s outputs in a 14 by 14 grid of patches, the entire image is represented by a tensor of shape $14 \times 14 \times 768$. This tensor is then processed by a decoder that upsamples the tensor via two deconvolutional layers resulting in a pose prediction tensor of shape $56 \times 56 \times 14$, where 14 is the number of channels (one for each joint). The model architecture is shown in Fig. 3. Before finetuning ViTPose on annotated pose estimation data, the ViTPose encoder is initialized from a pretrained ViT encoder. In ViTPose, they initialize with MAE’s encoder; we use this as a baseline to compare with ViTPose models initialized with our three specialized encoders. The base ViTPose outperformed TransPose and HRNet on the COCO dataset. [39]

VI. RESULTS AND DISCUSSION

A. Best Performing Pose Estimation Architecture on SMAIL Is ViTPose

The performance of the pose estimation methods and their variants are shown in Table II. The best-performing model was ViTPose, with a PCK of 84.57% and an NME of 17.9 mm, which is consistent with ViTPose’s claim as representing the state-of-the-art (SOTA). The next top-performing method
was HRNet-W48, with a PCK of 84.19% and an NME of 18.6 mm. Although the improvement in PCK was marginal, there was a significant increase in the NME, demonstrating that the ViTPose architecture has better localization than HRNet. The TransPose variants had the lowest overall performance of the baseline methods. The ResNet variant, TransPose-R-A4, had the lowest performance. TransPose-H-A4, which uses the HRNet-W32 backbone, achieved a lower PCK than HRNet-W32 alone; similarly, TransPose-H-A6, which uses the HRNet-W48 backbone, achieved a lower PCK than HRNet-W48 alone. Although TransPose was shown to achieve better performance on the COCO dataset when compared to HRNet, this did not hold true when training on the SMaL dataset. This may be due to the fact that TransPose only uses 5%–25% of the original backbones, significantly reducing the knowledge transfer of the pretrained backbones. Given the relatively small dataset of the depth and PSM modalities, the advantage of larger pretrained backbones is evident. The larger variants of the baseline methods consistently outperform the smaller models by significant margins. It is important to note that although the ViTPose model has the most parameters, it achieves the fastest inference speeds as demonstrated by its throughput (relative to the other models). It is roughly 50% faster than HRNet-W48 and 300% faster than Transpose-H-A6, the second- and third-best performing model architectures, respectively.

**B. Pretraining the ViT Backbone on Depth and Pressure Imagery Results in Increased Performance on SMaL**

Given that the ViT backbone is amenable to hierarchical MAE pretraining strategy, we can further improve the performance of the best-performing architecture, ViTPose. Evident from the results shown in Table III, there is a clear benefit to pretraining the ViT backbone on data that is similar to the target dataset. Even though the pretraining data was solely adult data and the majority synthetic, it remained transferable to the SMaL dataset. Introducing a significant amount of unlabeled depth and PSM data to the model successfully reduced the domain shift between the original ImageNet training set and the target SMaL dataset. The best-performing pretrained model was ViTPose-S-150 with a PCK of 86.06% and an NME of 16.9 mm. The addition of 50 epochs of pretraining helped to increase performance over the ViT-S-100. This was the overall best-performing model on both metrics and showed a statistically significant difference in performance when compared to ViT-Base ($p < 0.008$, where 0.05/6 since six pretraining variations were tested). It is important to note that when observing the overall performance across each fold, there is a natural variance between the fold. Since each fold is only made of 60 unique poses, a few difficult cases within the fold can cause performance drops across all models. Therefore, to test for significance, a paired t-test was used.

Interestingly, using SLP real data alone for pretraining resulted in worse performance than no pretraining; both ViT-R-100 and ViT-R-150 performed worse than ViT-Base. This may be due to the relatively small dataset size of SLP, resulting in overfitting on the MAE task. Using both real and synthetic datasets for pretraining showed increased performance compared to synthetic only, as ViT-B-100 outperformed ViT-S-100 and was the second-best pretraining strategy. ViT-B-100 is equivalent to ViT-S-150 with continued training on the SLP dataset for another 100 epochs. However, this was not the case for ViT-B-150, which had worse performance than ViT-S-150. This indicates that continued training on SLP for 100 epochs has a positive effect, but if this is increased by 50 epochs, it has a negative effect. A potential improvement to the pretraining strategy could be training on synthetic data for 150 epochs and continuing training on real data for only 100 epochs, to avoid overfitting on the SLP data. Overall, these results show the benefit of hierarchical pretraining for

### Table II

**Comparison of Pose Estimation Models on SMaL Dataset**

| Model       | Parameters | Throughput | Fold Accuracy | Average | PCK  | NME  |
|-------------|------------|------------|---------------|---------|------|------|
|             |            |            | 1  | 2  | 3  | 4  | 5  |       |       |
| HRNet-W32   | 28.5M      | 2.9x       | 82.80         | 76.25   | 83.27 | 85.42 | 79.76 | 81.50 | 20.4  |
| HRNet-W48   | 63.6M      | 2.1x       | **84.17**     | 80.00   | 85.42 | 87.74 | **83.63** | 84.19 | 18.6  |
| Transpose-R-A4 | 6.0M     | 2.5x       | 77.92         | 73.87   | 78.87 | 81.85 | 78.33 | 78.17 | 23.2  |
| Transpose-H-A4 | 17.3M   | 1.4x       | 81.19         | 76.90   | 82.44 | 82.68 | 80.71 | 80.79 | 22.1  |
| Transpose-H-A6 | 17.5M   | 1.0x       | 81.96         | 79.35   | 83.99 | 83.99 | 80.83 | 82.02 | 20.7  |
| ViTPose-Base | 90.0M      | 3.1x       | 83.51         | **81.96** | **86.61** | **87.86** | 82.92 | **84.57** | **17.9** |

### Table III

**Comparison of ViTPose Under Different Pretraining Conditions (* Denotes Significant Difference)**

| Model       | Pretraining | Epochs | Fold Accuracy | Average | PCK  | NME  |
|-------------|-------------|--------|---------------|---------|------|------|
|             |             |        | 1  | 2  | 3  | 4  | 5  |       |       |
| ViTPose-Base | -           | -      | 83.51         | 81.96   | 86.61 | 87.86 | 82.92 | 84.57 | 17.9  |
| ViTPose-S-100 | ✓          | -      | 83.75         | 82.20   | 86.73 | 87.20 | 84.58 | 84.89 | 17.6  |
| ViTPose-R-100 | ✓          | -      | 83.57         | 80.71   | 85.71 | 86.79 | 84.11 | 84.18 | 18.8  |
| ViTPose-B-100 | ✓          | 100    | 84.76         | 82.44   | 86.67 | **89.64** | **85.89** | 85.88 | 17.2  |
| ViTPose-S-150 | ✓          | 150    | **85.77**     | **83.21** | **87.92** | **88.75** | **84.64** | **86.06** | **16.9** |
| ViTPose-R-150 | ✓          | 150    | 83.57         | 81.49   | 85.83 | 87.44 | 84.17 | 84.80 | 18.3  |
| ViTPose-B-150 | ✓          | 150    | 85.30         | 81.73   | 86.61 | 87.62 | 83.87 | 85.02 | 17.6  |
under-the-cover pose estimation when there is a large enough dataset and caution is used to avoid overfitting on smaller datasets.

C. Under-the-Cover Infant Pose Can Be Accurately Inferred Using a Pose Estimation Model

Fig. 6 shows various test cases from SMaL and the best-performing method, ViTPose-S-150’s predictions. The accuracy of the pose predictions is impressive given the heavy occlusions present in the depth image and the low resolution of the pressure image. When compared to the performance on other datasets, our performance is comparable. Wu et al. [33] achieve a root-mean-square distance on all joints of 4.12 pixels on their uncovered depth-based infant dataset. When converting our best performance of 16.9 mm to pixels, we achieve an error of 3.33 pixels. The model gets the majority of keypoints correct, with most of the error on joints in the extremity. The slight errors on the wrist and ankle are observed in the first and third rows of Fig. 6, respectively. When breaking down the performance by joint type, as shown in Tables IV and V, it is clear that the wrist poses the most difficult challenge to the model, with the highest NME error and the lowest PCK. The hip and shoulder keypoints boast the highest accuracy of the joints. This is relevant to one of our motivations, under the cover of ROI detection, as the bounding box connecting these joints highlights the most important regions for noncontact RR estimation [5]. Tables IV and V also show the added difficulty when using thicker blanket covering, as the overall performance was worse under thick blankets, although not by a large margin. The third and fourth rows of Fig. 6 show an example of the thick cover condition. The fourth row presents an additional challenge, with the pose being very tight and the limbs close together. In spite of this, the majority of the prediction was correct.

D. Transferring Knowledge From Adult Pose to Infant Pose Results in Increased Performance

The two-stage training strategy is highly effective in improving performance on the SMaL dataset, as shown in Table VI. When excluding the SLP training stage, the performance drops by 12% and 9% for ViTPose-Base and ViTPose-S-150, respectively. This finding supports the conclusions of our previous work where adult pressure data was leveraged for transfer learning to infant pressure data [17]. The benefit of transferring knowledge from adult data is still apparent even when no labeled data is used. This is seen from the large performance gains after pretraining on unlabeled adult data, with a 9% increase of ViTPose-S-150 over ViTPose-Base after only training on SMaL data. Although transfer learning from adult data improves performance, adult data on its own is not sufficient. When training solely on the SLP dataset, we observed an average PCK of only 2.22%. With the abundance of available adult data and the relative lack of infant data, this transfer learning strategy can be applied to any task involving the domain shift between adults and infants.

E. Combining Depth and PSM Inputs Results in Increased Performance

When limiting the modality input to only depth or only PSM, we see significant drops in performance, as shown in Table VI. The combination of these modalities contributes more than a 10% increase in performance, pointing to the complete nature of the depth and pressure images. This result is in line with the results demonstrated by Liu et al. [12]. In all cases, the PSM-only training outperforms the depth-only, showing that individually the PSM is more informative than depth for the SMaL dataset. When the models are not trained on SLP data first (also using a single modality), we see larger performance drops when using depth alone compared to PSM alone. This shows that there is a higher variance in the depth data compared to the PSM data; however, this can be mitigated by first training on a larger depth-only dataset before.

F. Limitations of Mannequin Dataset

Although we have observed that it is possible to achieve accurate pose estimation for infants from under-the-cover
infant pose estimation, we have limited the study to a simulated environment using a mannequin. A real NICU environment introduces new challenges, including the bundling of patient limbs, the use of multiple layers of blankets, and the presence of extraneous items in the patient environment. In addition, the material properties of the

![Fig. 6. Predictions on a test set from the best performing model, ViTPose-S-150, along with ground-truth data.](image-url)
mannequin differ from real infants, whose bodies are more compliant.

Given these challenges, there is a need for a real-world dataset. However, obtaining ground-truth pose in a real dataset would be challenging. Nursing staff would need to temporarily remove blankets from the patient, capture an uncovered image for pose annotation, and then re-cover the patient before their pose changes, which would be disruptive to the routine care of the patient. To carry out such a study, we suggest alternative methods for collecting ground truth, such as using a blanket material that is transparent to allow annotation of pose, but not affect the depth image. More work would be required to determine the feasibility of this solution.

VII. CONCLUSION

In summary, we present and release a novel multimodal dataset, SMaL, for under-the-cover infant pose estimation. We use this dataset to train state-of-the-art pose estimation methods and evaluate their performance. We also leverage existing multimodal adult pose datasets for transfer learning. By implementing a transformer-based pose estimation method, we demonstrate a hierarchical pretraining and finetuning strategy using adult data to significantly boost performance on our SMaL dataset. Overall, we were able to accurately infer the infant pose in the presence of heavy occlusion from blankets using depth and pressure imagery.

REFERENCES

[1] P. Hamilton, M. Curley, and R. Aim, “Effect of adaptive motion-artifact reduction on QRS detection,” Biomed. Instrum. Technol., vol. 34, no. 3, pp. 197–202, 2000.
[2] M. Wolf et al., “Improved monitoring of preterm infants by fuzzy logic,” Technol. Health Care, Off. J. Eur. Soc. Eng. Med., vol. 4, no. 2, pp. 193–201, Aug. 1996.
[3] C. Lund, L. B. Nonato, J. M. Kuller, L. S. Franck, C. Cullander, and D. K. Durand, “Disruption of barrier function in neonatal skin associated with adhesive removal,” J. Pediatr., vol. 131, no. 3, pp. 367–372, Sep. 1997.
[4] D. G. Kyrollos, K. Greenwood, J. Harrold, and J. R. Green, “Detection of false alarms in the NICU using pressure sensitive mat,” in Proc. IEEE Sensors Symp. (SAS), Aug. 2021, pp. 1–5.
[5] L. Maurya, P. Kaur, D. Chawla, and P. Mahapatra, “Non-contact breathing rate monitoring in newborns: A review,” Comput. Biol. Med., vol. 132, May 2021, Art. no. 104321.
[6] D. G. Kyrollos, J. B. Tanner, K. Greenwood, J. Harrold, and J. R. Green, “Noncontact neonatal respiration rate estimation using machine vision,” in Proc. IEEE Sensors Appl. Symp. (SAS), Aug. 2021, pp. 1–6.
[7] S. Nizami, A. Bekele, M. Hozayen, K. J. Greenwood, J. Harrold, and J. R. Green, “Measuring uncertainty during respiratory rate estimation using pressure-sensitive mats,” IEEE Trans. Instrum. Meas., vol. 67, no. 7, pp. 1535–1542, Jul. 2018.
[8] AAP Task Force on Infant Positioning and SIDS, “Positioning and SIDS,” Pediatrics, vol. 89, no. 6, pp. 1120–1126, Jun. 1992.
[9] S. Bembich, C. Oretti, L. Travan, A. Clarici, S. Massacesi, and S. Demarini, “Effects of prone and supine position on cerebral blood flow in preterm infants,” J. Pediatrics, vol. 160, no. 1, pp. 162–164, Jan. 2012.
[10] P. Fister, M. Nolimal, H. Lenasi, and M. Klemenc, “The effect of sleeping position on heart rate variability in newborns,” BMC Pediatrics, vol. 20, no. 1, pp. 1–8, Apr. 2020.
[11] H.-L. Kukkola and T. Kirjavainen, “Obstructive sleep apnea is position dependent in young infants,” Pediatric Res., pp. 1–7, Aug. 2022.
[12] S. Liu, X. Huang, N. Fu, C. Li, Z. Su, and S. Ostadabbas, “Simultaneously-collected multimodal lying pose dataset: Enabling in-bed human pose monitoring,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 45, no. 1, pp. 1106–1118, Jan. 2023.
[13] D. Groos, L. Adde, R. Stoen, H. Ramampiaro, and E. A. F. Ihlen, “Towards human-level performance on automatic pose estimation of infant spontaneous movements,” Comput. Med. Imag. Graph., vol. 95, Jan. 2022, Art. no. 102012.
[14] X. Huang, N. Fu, S. Liu, and S. Ostadabbas, “Invariant representation learning for infant pose estimation with small data,” in Proc. 16th IEEE Int. Conf. Autom. Face Gesture Recognit. (FG), Dec. 2021, pp. 1–8.
[15] H. M. Clever, P. L. Grady, G. Turk, and C. C. Kemp, “BodyPressure—Inferring body pose and contact pressure from a depth image,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 45, no. 5, pp. 137–153, Jan. 2023.
[16] C. J. Reed et al., “Self-supervised pretraining improves self-supervised pretraining,” in Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis. (WACV), Jan. 2022, pp. 2584–2594.
[17] D. G. Kyrollos, K. Greenwood, J. Harrold, and J. R. Green, “Transfer learning approaches for neonate head localization from pressure images,” in Proc. IEEE Int. Symp. Med. Meas. Appl. (MeMeA), Jun. 2022, pp. 1–6.
[18] S. Liu and S. Ostadabbas, “A vision-based system for in-bed posture tracking,” in Proc. IEEE Int. Conf. Comput. Vis. Workshops (ICCVW), Oct. 2017, pp. 1373–1382.
[19] D. G. Kyrollos, R. Hassan, Y. S. Dosso, and J. R. Green, “Fusing pressure-sensitive mat data with video through multi-modal registration,” in Proc. IEEE Int. Symp. Med. Meas. Technol. Conf. (IEMTc), May 2021, pp. 1–6.
[20] H. M. Clever, A. Kapusta, D. Park, Z. Erickson, Y. Chitalia, and C. C. Kemp, “3D human pose estimation on a configurable bed from a pressure image,” in Proc. IEEE/RSJ Int. Conf. Intel. Robots Syst. (IROS), Oct. 2018, pp. 54–61.
[21] L. Casas, N. Navab, and S. Demirci, “Patient 3D body pose estimation from pressure imaging,” Int. J. Comput. Assist. Radiol. Surg., vol. 14, no. 3, pp. 517–524, Dec. 2018.
[22] D. G. Kyrollos, K. Greenwood, J. Harrold, and J. R. Green, “Transfer learning approaches for neonate head localization from pressure images,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2020, pp. 6215–6224.
[23] T. Dayarathna et al., “Privacy-preserving in-bed pose monitoring: A fusion and reconstruction study,” Expert Syst. Appl., vol. 213, Mar. 2023, pp. 119139.
[24] Y. Yin, J. P. Robinson, and Y. Fu, “Multimodal in-bed pose and shape estimation under the blankets,” in Proc. 30th ACM Int. Conf. Multimedia, Oct. 2022, pp. 2411–2419.
[25] M. Afham, U. Haputhanthri, J. Pradeepkumar, M. Anandakumar, A. De Silva, and C. U. S. Edussooryira, “Towards accurate cross-domain in-bed human pose estimation,” in Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP), Jan. 2021, pp. 3965–3969.
[26] H. M. Clever, A. Kapusta, D. Park, Z. Erickson, Y. Chitalia, and C. C. Kemp, “BodyPressure—an automatic in-bed pose estimation system,” in Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS), Oct. 2018, pp. 54–61.
[27] L. Casas, N. Navab, and S. Demirci, “Patient 3D body pose estimation from pressure imaging,” Int. J. Comput. Assist. Radiol. Surg., vol. 14, no. 3, pp. 517–524, Dec. 2018.
[28] D. G. Kyrollos, K. Greenwood, J. Harrold, and J. R. Green, “Fusing pressure-sensitive mat data with video through multi-modal registration,” in Proc. IEEE Int. Symp. Med. Meas. Technol. Conf. (IEMTc), May 2021, pp. 1–6.
[29] T. Cao, M. A. Armin, S. Deman, L. Pettersson, and D. Ahmedt-Aristizabal, “In-bed human pose estimation from unseen and privacy-preserving image domains,” in Proc. IEEE 19th Int. Symp. Biomed. Imag. (ISBI), Mar. 2022, pp. 1–5.
[30] T. Dayaratha et al., “Privacy-preserving in-bed pose monitoring: A fusion and reconstruction study,” Expert Syst. Appl., vol. 213, Mar. 2023, pp. 119139.

TABLE VI

| Model    | SLP + SMaL | SMaL Only |
|----------|------------|-----------|
|          | Both Depth | PSM       | Both Depth | PSM       |
| ViTPost-Base | 64.87 13.30 | 74.77 72.48 | 64.27 66.49 |
| ViTPost-3-130 | 86.06 73.56 | 75.02 77.95 | 57.02 64.81 |


[32] A. Rihar, M. Mihelj, J. Palič, J. Kolar, and M. Munih, “Infant trunk posture and arm movement assessment using pressure mattress, inertial and magnetic measurement units (IMUs),” J. NeuroEng. Rehabil., vol. 11, no. 1, p. 133, 2014.
[33] Q. Wu et al., “Supine infant pose estimation via single depth image,” IEEE Trans. Instrum. Meas., vol. 71, pp. 1–11, 2022.
[34] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick, “Masked autoencoders are scalable vision learners,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2022, pp. 16000–16009.
[35] A. Vaswani et al., “Attention is all you need,” in Proc. Adv. neural Inf. Process. Syst., vol. 30, 2017, pp. 1–11.
[36] A. Dosovitskiy et al., “An image is worth 16 × 16 words: Transformers for image recognition at scale,” 2020, arXiv:2010.11929.
[37] Z. Peng, L. Dong, H. Bao, Q. Ye, and F. Wei, “BEiT v2: Masked image modeling with vector-quantized visual tokenizers,” 2022, arXiv:2208.06366.
[38] Y. Li, H. Mao, R. Girshick, and K. He, “Exploring plain vision transformer backbones for object detection,” 2022, arXiv:2203.16527.
[39] Y. Xu, J. Zhang, Q. Zhang, and D. Tao, “ViTPose: Simple vision transformer baselines for human pose estimation,” 2022, arXiv:2204.12484.
[40] S. Abnar, M. Dehghani, B. Neyshabur, and H. Sedghi, “Exploring the limits of large scale pre-training,” 2021, arXiv:2110.02095.
[41] K. Sun, B. Xiao, D. Liu, and J. Wang, “Deep high-resolution representation learning for human pose estimation,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2019, pp. 5693–5703.
[42] S. Yang, Z. Quan, M. Nie, and W. Yang, “TransPose: Keypoint localization via transformer,” in Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV), Oct. 2021, pp. 11802–11812.

Daniel G. Kyrollos (Student Member, IEEE) received the B.Eng. degree in biomedical and electrical engineering and the M.A.Sc. degree in electrical and computer engineering with a specialization in data science from Carleton University, Ottawa, Canada, in 2020 and 2022, respectively. His thesis project investigated novel patient monitoring technologies in the neonatal intensive care unit (NICU). His research interests include data science, machine learning, machine vision, natural language processing, and signal processing.

Anthony Fuller (Student Member, IEEE) received the B.Eng. degree in aerospace engineering from Carleton University, Ottawa, ON, Canada, in 2015, where he is currently pursuing the M.A.Sc. degree in electrical and computer engineering.

His research interests include self-supervised representation learning for various applications, such as remote sensing and patient monitoring.

Kim Greenwood (Senior Member, IEEE) received the Diploma degree in electrical engineering technology from the Ryerson Polytechnical Institute, Toronto, ON, Canada, in 1984, the B.A.Sc. degree in technology management from Bemidji State University, Bemidji, MN, USA, in 2006, and the M.A.Sc. degree in biomedical engineering from Carleton University, Ottawa, ON, Canada, in 2010.

He was recognized as a fellow of The Engineering Institute of Canada, Ottawa, ON, Canada, in 2018, and as a fellow of the Canadian Medical and Biological Engineering Society, Ottawa, ON, Canada, in 2020. He is a Licensed Professional Engineer with the Province of Ontario and is a Certified Clinical Engineer. He is an Adjunct Professor in mechanical engineering with the University of Ottawa, Ottawa, the Executive Director of the Eastern Ontario Clinical Engineering Service, and the Chief Clinical Equipment Officer at the Children’s Hospital of Eastern Ontario, Ottawa.

JoAnn Harrold received the B.Sc. and M.D. degrees from McMaster University, Hamilton, ON, Canada, in 1994 and 1997, respectively. She holds a Royal College certification in Pediatrics (training at the University of Toronto, Toronto, ON, Canada) and in Neonatal-Perinatal Medicine (training at McMaster University).

She is currently an Associate Professor with the Faculty of Medicine, University of Ottawa, Ottawa, ON, Canada, the Division Chief of Neonatology with the Children’s Hospital of Eastern Ontario, Ottawa, and the Division Head of Newborn Care with The Ottawa Hospital, Ottawa.

James R. Green received the B.A.Sc. degree in systems design engineering from the University of Waterloo, Waterloo, ON, Canada, in 1998, the M.A.Sc. and Ph.D. degrees in electrical and computer engineering from Queen’s University, Kingston, ON, Canada, in 2000 and 2005, respectively.

He is currently a Professor with the Department of Systems and Computer Engineering, Carleton University, Ottawa, ON, Canada. His research interests include machine-learning challenges within biomedical informatics, patient monitoring, computational acceleration of scientific computing, and the design of novel assistive devices.