

# Leveraging Synthetic Adult Datasets for Unsupervised Infant Pose Estimation

Sarosij Bose, Hannah Dela Cruz, Arindam Dutta, Elena Kokkoni,  
 Konstantinos Karydis, Amit K. Roy-Chowdhury  
 University of California, Riverside, USA

{sbose007, hdela004, adutt020, elenak}@ucr.edu, {kkarydis, amitrc}@ece.ucr.edu

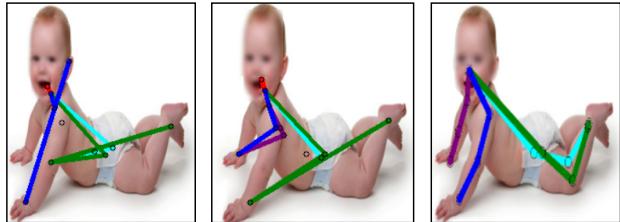
*This paper has been accepted at 8<sup>th</sup> Workshop and Competition on Affective & Behavior Analysis in-the-wild (ABAW) held in conjunction with the IEEE Computer Vision and Pattern Recognition Conference (CVPR) 2025.*

## Abstract

*Human pose estimation is a critical tool across a variety of healthcare applications. Despite significant progress in pose estimation algorithms targeting adults, such developments for infants remain limited. Existing algorithms for infant pose estimation, despite achieving commendable performance, depend on fully supervised approaches that require large amounts of labeled data. These algorithms also struggle with poor generalizability under distribution shifts. To address these challenges, we introduce SHIFT: Leveraging SyntHetic Adult Datasets for Unsupervised InFanT Pose Estimation, which leverages the pseudo-labeling-based Mean-Teacher framework to compensate for the lack of labeled data and addresses distribution shifts by enforcing consistency between the student and the teacher pseudo-labels. Additionally, to penalize implausible predictions obtained from the mean-teacher framework we also incorporate an infant manifold pose prior. To enhance SHIFT’s self-occlusion perception ability, we propose a novel visibility consistency module for improved alignment of the predicted poses with the original image. Extensive experiments on multiple benchmarks show that SHIFT significantly outperforms existing state-of-the-art unsupervised domain adaptation (UDA) based pose estimation methods by  $\sim 5\%$  and supervised infant pose estimation methods by a margin of  $\sim 16\%$ . The project page is available at [sarosijbose.github.io/SHIFT](https://sarosijbose.github.io/SHIFT).*

## 1. Introduction

Estimating pose keypoints in infants is a challenging task with several biomedical applications. Key examples in-



**Figure 1. Need for unsupervised domain adaptive infant pose estimation.** From left to right keypoint predictions from a baseline adult human pose estimation model [20], predictions from a SOTA UDA pose estimation model [14], and predictions from our method, SHIFT. Adult pose estimation models fail when directly applied to infant data; similarly, UniFrame [20] struggles to overcome the domain shift between adults and infants. In contrast, SHIFT accounts for the highly self-occluded pose distribution of infants, thereby effectively adapting to the infant domain.

clude neuromotor assessment in infants at risk for developmental disorders [35, 44], facial expression identification [27, 43], safety monitoring [9], as well as feedback control design in wearable assistive robotics for this population [28, 29]. Existing algorithms for infant pose estimation [11, 14, 51] predominantly rely on fully-supervised training to achieve state-of-the-art results on curated datasets [11, 14]. However, these datasets typically contain limited data points captured in relatively controlled settings. As a result, models trained on these existing datasets often overfit them and struggle to generalize on out-of-domain images. Moreover, privacy and ethical concerns related to the use of infant data, coupled with the labor-intensive, time-consuming, and challenging task of annotating infant poses, limit the development of effective large-scale infant pose estimation datasets [25]. In contrast, there is an abundance of publicly available adult human pose estimation datasets (e.g., [17, 42]). This motivates us to ask: *Is it possible to adapt a pre-trained adult pose estimation model to the task of infant pose estimation in an unsupervised setting?*

Recent studies [18, 20] have proposed methods for adapting adult pose estimation models trained on synthetic source data to in-the-wild target images. However, these

methods perform sub-optimally when the target dataset consists exclusively of infant pose images. Plausible causes for this include key anatomical differences between adults and infants [15], variations in movement patterns that become adult-like in later years [22, 38] as well as the more varied body poses attained by infants as compared to adults [36], which are not typically captured in the adult datasets employed in such adaptation-based pose estimation algorithms [14]. Recent works such as Huang *et al.* [14] have attempted to rectify these issues by fine-tuning a model trained on synthetic adult datasets [42] to infant datasets in a fully-supervised manner. However, this approach requires access to labeled infant pose datasets and results in poor generalizability across other infant datasets as illustrated in Figure 1.

To address these challenges, we develop an unsupervised domain adaptation (UDA) algorithm termed SHIFT: *Leveraging SyntHetic Adult Datasets for Unsupervised InFanT Pose Estimation*. SHIFT incorporates the mean-teacher model training methodology [39], which updates the teacher model’s weights with an exponential moving average of the student model’s weights. This approach ensures the generation of reliable pseudo-labels to guide the adaptation process, compensating for the lack of target domain ground truth labels. We leverage the data augmentation principle to enforce self-supervised consistency between the student model’s predictions and those of the teacher model [20]. Compared to directly using the pre-trained adult pose estimation model, enforcing consistency in the feature space improves adaptation performance. However, consistency enforcement alone does not address the prediction of physically implausible poses caused by insufficient anatomical understanding of infants.

We thus incorporate two novel regularizers that introduce anatomical constraints specific to infants, aiding the adapted model in predicting accurate infant poses. First, we train an infant-specific parametric pose prior that is inspired by [40] during the adaptation process. Leveraging the principles of manifold hypothesis [5], we design the prior such that anatomically plausible poses exist as manifold points on a zero-level set [5], while physically implausible poses lie at a non-zero distance from the manifold. This infant pose prior provides plausibility regularization during the adaptation phase by penalizing the model for anatomically implausible predictions. However, this proposed pose prior does not include contextual information from the image itself. As a result, the model may predict anatomically plausible poses that do not align with the infant’s pose in the RGB image, especially under significant self-occlusions. To address this, the second regularization technique enforces self-supervised consistency between the predicted keypoints and the segmentation mask of the infant. This is achieved by training a function that learns to map a given set of pose keypoints to a silhouette and a pre-

trained segmentation model [2] that extracts the segmentation masks of the infant in target images.

In summary, our **main contributions** are:

- We propose SHIFT, a novel Unsupervised Domain Adaptation (UDA) framework to adapt a pre-trained 2D adult pose estimation model to infants. To the best of our knowledge, this is the first UDA based work to address infant pose estimation.
- In addition to leveraging feature consistency, SHIFT employs an infant-specific manifold pose prior, trained offline to capture physically plausible infant poses. To address high self-occlusion, we incorporate additional context to ensure pose-image consistency.
- We conduct extensive qualitative and quantitative evaluations on two challenging infant pose datasets, demonstrating that our method significantly outperforms ( $\approx 5\%$ ) existing analogous UDA methods, as well as, outperforms supervised infant pose estimation methods by  $\approx 16\%$ .

## 2. Related Works

**Human Pose Estimation.** Human pose estimation involves the localization of anatomical joints on the human body, such as the head, shoulders and knees. Existing algorithms for this task can be primarily categorized into two paradigms: bottom-up methods and top-down methods. Top-down methods, which require a detection step before pose estimation, are often more accurate than bottom-up methods. HourGlass [30] was one of the first proposed top-down algorithms, relied on the regression of 2D Gaussian heatmaps to individual keypoints. Since then, several other top-down approaches [4, 19, 37, 45, 47, 49] have been developed. In contrast, bottom-up algorithms [1, 3, 7, 16, 31, 33] estimate all possible keypoints in an image and then perform a data association step to assign keypoints to individuals. Notably, these methods require extensively annotated datasets, making them less scalable for scenarios where there are limited or no annotations.

**Infant Pose Estimation.** Infant pose estimation is a subset of human pose estimation that specifically targets localizing keypoints for infants. Hesse *et al.* [11] introduced the benchmark MINI-RGBD dataset by utilizing the statistical 3D shape model *Skinned-Multi Infant Linear (SMIL)* to generate synthetically masked RGB video sequences of real infants in motion. Building on [11], Huang *et al.* [14] proposed the SyRIP dataset, which contains both real and synthetic infants, the latter being generated by fitting the SMIL [11] model. ZEDO-i [51] performs a 2D-to-3D lifting operation of ground truth infant poses using a Score-Matching Network (SMN) which is driven by a conditional diffusion model. These works rely on ground truth labels in the infant domain; SHIFT eliminates this dependency by addressing the problem of 2D keypoint

estimation in infants in an unsupervised manner.

**UDA for Pose Estimation.** UDA algorithms (e.g., [8, 12, 24, 32, 46, 50]) aim to transfer knowledge from a model trained on a labeled source domain to an unlabeled target dataset), removing the need for target domain annotations. Recent work in pose estimation [18] proposed the use of adversarial training to learn domain-invariant features, facilitating the transfer of knowledge from the labeled source domain to the unlabeled target domain. Kim *et al.* [20] employed the mean-teacher framework [39] and the style-transfer technique [13] to refine pseudo-labels on the unlabeled target data, thus facilitating both output-level and input-level alignments respectively and achieving state-of-the-art results. Inspired by [20], we propose a novel algorithm for infant pose estimation by transferring knowledge from a labeled adult human dataset.

### 3. Methodology

In the source domain  $S$ , we have a labeled adult pose dataset  $\mathcal{D}_S = \{(x_s^i, y_s^i)\}_{i=1}^{N_s}$  which consists of  $N_s$  images,  $x_s \in \mathbb{R}^{H \times W \times 3}$  where  $H$  and  $W$  refer to the spatial dimensions of the image, and the corresponding ground truth 2D keypoints  $y_s \in \mathbb{R}^{K \times 2}$  where  $K$  represents the respective coordinates of the keypoints. In the target domain  $T$ , we have an unlabeled infant pose dataset  $\mathcal{D}_T = \{x_t^i\}_{i=1}^{N_t}$  comprised of  $N_t$  images,  $x_t \in \mathbb{R}^{H \times W \times 3}$ . A 2D pose estimation model  $\mathcal{M}$  is pretrained on the source domain  $S$  to predict keypoint heatmaps. We seek to adapt this model  $\mathcal{M}$  to the unlabeled target domain  $T$  to achieve improved performance in comparison to the unadapted source model. An overview of the framework is shown in Figure 2. In the following sections, we describe the different components of SHIFT as follows,

- In Section 3.1, we describe the pre-training methodology of the pose estimation model  $\mathcal{M}$  to provide a weight initialization on the source domain  $S$ .
- In Section 3.2, we describe the adaptation process in the estimation space using the mean-teacher paradigm and show how it is insufficient to capture the anatomical and semantic information for infants.
- In Section 3.3, we describe the working principle of our manifold pose prior and outline its design to capture the intricate anatomical details from infant poses.
- In Section 3.4, we elaborate on the pose-image consistency module and describe how it performs the adaptation by aligning the visibility between the predicted poses and segmentation masks of infants.

#### 3.1. Source Domain Pre-Training

We employ the source domain pretraining approach to initialize the pose estimation model  $\mathcal{M}$ . Following [41], we generate 2D Gaussian heatmaps, given by a conversion function  $\phi : \mathbb{R}^{K \times 2} \rightarrow \mathbb{R}^{K \times H' \times W'}$  where  $H'$  and  $W'$

refer to the spatial dimensions of the obtained heatmap. We then pass the source domain images  $x_s$  to generate source Gaussian heatmaps  $H_s$ , which can be represented as  $H_s = \phi(y_s)$ . The source (adult) domain pretraining is carried out using the MSE loss in a supervised fashion, i.e.

$$\mathcal{L}_{\text{sup}} = \frac{1}{N_s} \sum_{x_s \in D_s} \|\phi(\mathcal{M}(x_s)) - H_s\|_2 . \quad (1)$$

#### 3.2. Estimation Space Adaptation

It has been shown that weight-averaged training steps are better for model training rather than just the final model [26]. Therefore, similar to the mean-teacher [39] setup, we have a pretrained student model  $\mathcal{M}_s$  and a teacher model  $\mathcal{M}_t$ . At time  $t=0$ , both the weight-initialized student model  $\mathcal{M}_s$  and the teacher model  $\mathcal{M}_t$  are updated by the exponential moving average (EMA) of the student model's weights ( $\theta_{\mathcal{M}_s}$ ) to the teacher model's weights ( $\theta_{\mathcal{M}_t}$ ) as

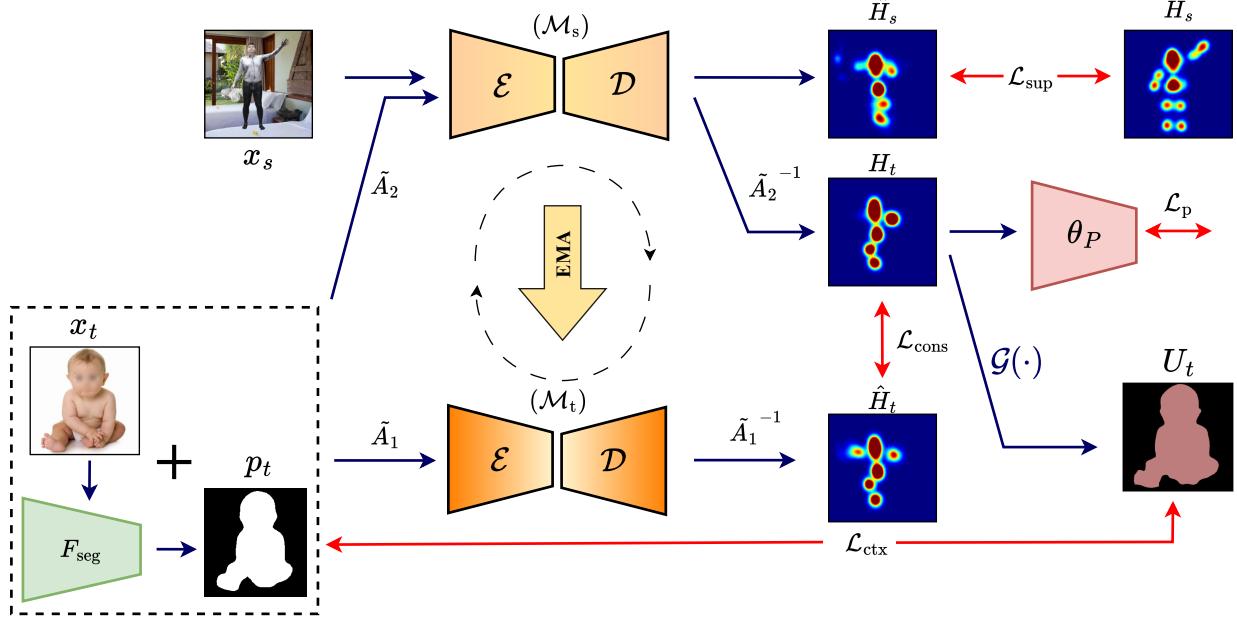
$$\theta_{\mathcal{M}_t} = \alpha \theta_{\mathcal{M}_{t-1}} + (1 - \alpha) \theta_{\mathcal{M}_s} . \quad (2)$$

The decay rate  $\alpha$  is set to 0.999. This is done to balance the teacher model's weights between the previous parameters and the latest updates, ensuring that the teacher model  $\theta_{\mathcal{M}_t}$  is updated with the student  $\theta_{\mathcal{M}_s}$  to prevent catastrophic forgetting. This stabilization is crucial for pretraining, as it enhances the teacher model's ability to generate reliable pseudo-labels for the unlabeled target data by avoiding overfitting, thereby improving the overall training efficacy.

We generate two different views of incoming target images  $x_t$  by performing augmentations  $\tilde{A}_1$  and  $\tilde{A}_2$  to the inputs of the student model  $\mathcal{M}_s$  and the teacher model  $\mathcal{M}_t$ , respectively. Similar to [20], we selectively patch out keypoints for which the teacher model  $\mathcal{M}_t$  produces the highest activations given by a patching operation  $P$ :  $\hat{x}_t = P(\tilde{A}_2(x_t))$ . This helps steer the student model's heatmap predictions  $H_t = \phi(\mathcal{M}_s(\hat{x}_t))$  to focus more on those keypoints where its confidence is relatively lower with respect to occluded keypoints. To generate pseudo-labels from  $\mathcal{M}_t$ , we sample only those keypoints that produce the maximum activation  $\hat{y}_t = \text{argmax}(\hat{H}_t)$  where  $\hat{H}_t = \phi(\mathcal{M}_t(\tilde{A}_1(x_t)))$ . In addition, to reduce the effect of noisy label propagation, we set a fixed threshold  $\tau_c$  to filter out unreliable pseudo-labels. Thus, the learning objective for the student model  $\mathcal{M}_s$  for the  $k^{\text{th}}$  keypoint on the student heat map  $\hat{H}_t^k$  and the pseudo-label  $\hat{x}_t^k$  is defined by the MSE loss

$$\mathcal{L}_{\text{cons}} = \frac{1}{N'_t} \sum_{x_t \in D_T} \sum_{k=0}^K (\hat{H}_t^k \geq \tau_c) \|\tilde{A}_1^{-1}(\hat{H}_t^k) - \tilde{A}_2^{-1}(\mathcal{M}_s(\hat{x}_t^k))\|_2 . \quad (3)$$

$N'_t$  refers to the batch size of the incoming target (infant) domain images and  $\tilde{A}_1^{-1}$  and  $\tilde{A}_2^{-1}$  refers to the inverse



**Figure 2. Framework Overview.** SHIFT utilizes the Mean-Teacher framework [39] to update the teacher model  $\mathcal{M}_t$  with an Exponential Moving Average (EMA) of the student model  $\mathcal{M}_s$ 's weights to adapt the model pre-trained on a labeled adult source dataset ( $x_s, y_s$ ) to unlabeled infant target images ( $x_t$ ) (Section 3.2). To address anatomical variations in infants, SHIFT employs an infant pose prior  $\theta_p$  which assigns plausibility scores for each prediction of the student model  $\mathcal{M}_s$  (Section 3.3). Further, to handle the large self-occlusions in the target domain, we employ an off-the-model  $F_{seg}$  to give pseudo segmentation masks  $p_t$  with which our  $Kp2Seg$  module  $\mathcal{G}(\cdot)$  learns to perform pose-image visibility alignment (Section 3.4) hence effectively leveraging the context present in the visible portions of each image. All the learnable components of the framework are denoted in red and rest in black.

augmentations after passing through the teacher and student models respectively. The  $\tau_c$  helps to take into account only the highly confident pseudo-labels and reject the rest. Therefore, this loss is crucial in enforcing consistency between the pseudo-labels produced by the teacher model and the student model's estimated keypoints.

### 3.3. Manifold Infant Pose Prior

Enforcing consistency between the estimation spaces of the student model  $\mathcal{M}_s$  and the teacher model  $\mathcal{M}_t$  lacks awareness about the different anatomical aspects of the target domain. To address this, we design a manifold prior for modeling all physically plausible poses as a zero-level set. It leverages the architecture introduced in PoseNDF [40] to generate a score based on the physical plausibility of infant poses. This module is trained offline in a *cross-dataset manner*, i.e. the prior is pre-trained on a different dataset than the dataset on which the evaluation is being carried out. The resulting pose prior module,  $(\theta_p)$ , can assess poses generated by the student model against a diverse set of anatomical variations.

To design a domain-agnostic representation for poses, we first define a set of anatomically connected pairs of joints in the human body. These are represented in the form of 2D orientation vectors which are given by

$$\mathcal{V} = (\theta_1^i, \theta_2^i, \dots, \theta_L^i), \theta_l \in \mathbb{R}^2. \quad (4)$$

Leveraging the manifold hypothesis [5], we assume that all physically plausible infant poses can lie on a manifold defined as a zero-level set  $\mathcal{P} = \{\theta \in \mathcal{V} \mid p(\theta) = 0\}$ , where  $l$  gives us the distance at which a particular pose may lie from the manifold. We use an encoding function  $p_{enc}$  to store this chain of individual pose orientation encodings. Each encoding can be written as

$$v_1 = p_{enc}^1(\theta_1), \quad v_i = p_{enc}^l(\theta_l, v_f), \quad l \in \{2, \dots, L\}, \quad (5)$$

while the overall pose encoding is represented as

$$p = [v_1 \mid \dots \mid v_L]. \quad (6)$$

Following [40], we then construct a prior dataset of pose encoding-distance pairs  $D_{pd} = \{(\theta_i, l_i)\}_{i=1}^{N_t}$  for the prior model  $\theta_p$  to train on. All poses that lie on the manifold are assigned  $l = 0$ . These generated pose encodings are not paired with the RGB images, which saves storage since there is no need for storing paired RGB-Pose information, and is domain agnostic as well. The prior model is then trained using a multi-stage approach to ensure the robustness of the prior (see Section 4.1). We generate

noisy poses by sampling noise from the von Mises distribution [6]. Given the  $i^{th}$  target domain image  $x_t^i$ , we obtain the heatmap  $H_t^i = \phi(\mathcal{M}_s(x_t^i))$  and compute a set of normalized orientation vectors from the pixel coordinates of the obtained heatmaps  $H_t^i$  in the target domain. These normalized orientation vectors are calculated for each pair of anatomically connected keypoints in  $D_{pd}$ . Therefore, for a given pair of connected keypoints  $(m, n)$ , we have the unit vector  $\theta$  starting from the direction of the initial pixel coordinate  $p^m$  to  $p^n$ . This can be represented as

$$\theta_{(m,n)} = \frac{p^m - p^n}{\|p^m - p^n\|_2} \quad \forall (m, n) \in \mathcal{V}. \quad (7)$$

**Adaptation with Pose Prior.** During adaptation, we leverage this trained prior to predict an average plausibility score based on the distance between the predicted pose of the student model and the plausible pose manifold. We can set the objective of the prior ( $\theta_p$ ) to predict the distance ( $l$ ) between the predicted pose and our pre-learned set of plausible poses on the manifold. This is given by

$$\mathcal{L}_p = \frac{1}{N'_t} \sum_{x_t \in D_t} \theta_p(T(\mathcal{M}_s(x_t))), \quad (8)$$

where  $N'_t$  refers to the batch size of images in the target domain and  $T$  is a differentiable orientation function to convert the student model's predictions into a set of orientation vectors which the prior can then process to give an average plausibility score.

### 3.4. Context Aware Adaptation

Most infant pose datasets suffer from the problem of high self-occlusions, caused for example when infants are in lateral recumbent or prone positions. Motivated by works that show that parsing multiple modalities can serve as a rich source of spatial information [23, 48], we utilize segmentation masks to provide additional contextual guidance to the student model for aligning the pose and image spaces during training. To ensure that our framework can accurately estimate infant poses even in challenging scenarios, we first extract binary foreground-background segmentation masks from the target domain  $\mathbf{T}$  using DeepLab-v3 [2]. We denote this pre-extracted set of pseudo masks as  $\mathbf{D}_{seg} = \{p_t^i\}_{i=1}^{N_t}$ , given that there are  $N_t$  images in the target dataset  $\mathbf{D}_T$ . In the student model ( $\mathcal{M}_s$ )'s estimation space, we transform the obtained heatmaps into segmentation masks using our Kp2Seg module  $\mathcal{G}$  (more details in Section 3.4). Given the  $i^{th}$  target domain image  $x_t^i$ , we have  $\hat{H}_t^i = \phi(\mathcal{M}_s(\tilde{A}_2(x_t^i)))$  as the  $i^{th}$  estimated heatmap of the target batch from the student model, hence leading to the segmentation mask  $U_t^i = \mathcal{G}(\hat{H}_t^i)$ , where  $\mathcal{G} : \mathbb{R}^{K \times 2} \rightarrow \mathbb{R}^{H'' \times W''}$  denotes our mapping module to convert heatmaps to segmentation masks, with  $H''$  and  $W''$

representing the spatial dimensions of the obtained mask. The self-supervised consistency objective between  $p_j^i$  and  $U_t^i$  can be written using the Cross-Entropy Loss function

$$\mathcal{L}_{ctx} = \frac{1}{N'_t} \sum_{i=1}^{N'_t} \sum_{j=1}^{H'' \times W''} -p_j^i \log(U_{t,j}^i), \quad (9)$$

where  $N'_t$  is the batch size of images in the target domain,  $p_j^i$  is the binary label (either 0 or 1) at pixel  $j$  of the  $i^{th}$  pre-extracted pseudo mask, and  $U_{t,j}^i$  is the predicted probability at pixel  $j$  of the  $i^{th}$  mapped segmentation mask.

**Keypoint to Segmentation Mapping.** Our keypoint to segmentation encoder module ( $\mathcal{G}$ ), serves as a mapping function to convert the predicted student heatmaps ( $\hat{H}_s$ ) to segmentation masks  $U$ . This is a non-trivial operation as it involves mapping the sparse low-resolution heatmaps representing the keypoints, into dense high-resolution segmentation maps. We utilize the decoder from the DC-GAN [34] architecture which serves as a learned mapping function to convert heatmap predictions to segmentation masks. To train  $\mathcal{G}$ , we prepare a synthetic auxiliary set which comprises ground-truth poses and segmentation masks given by  $D_{aux} = \{(g_s^i, p_s^i)\}_{i=1}^{N_s}$ , where  $g_s^i$  and  $p_s^i$  refer to the  $i^{th}$  pose and pre-extracted segmentation masks from the auxiliary set  $D_{aux}$  respectively.  $\mathcal{G}$  is then trained in an end-to-end supervised manner to map segmentation masks from keypoints from the auxiliary set. This can be framed as a supervised objective between the pre-extracted segmentation masks  $p_s$  and the mapped segmentation masks  $u_s = \mathcal{G}(\phi(g_s^i))$  in terms of the Cross-Entropy Loss

$$\mathcal{L}_{\mathcal{G}} = \frac{1}{N_s} \sum_{i=1}^{N_s} \sum_{j=1}^{H'' \times W''} -p_{s,j}^i \log(u_{s,j}^i), \quad (10)$$

where  $N_s$  stands for the batch size of source images. Note that we do not employ ground truth segmentation masks either for the adaptation process or for offline training and we do not employ any RGB information for training  $\mathcal{G}$  so it is domain agnostic.

### 3.5. Overall Adaptation

Combining all the aforementioned losses, our student model  $\mathcal{M}_s$  is trained using the weighted adaptation objective

$$\mathcal{L}_{total} = \mathcal{L}_{sup} + \lambda_{cons} \mathcal{L}_{cons} + \lambda_p \mathcal{L}_p + \lambda_{ctx} \mathcal{L}_{ctx}. \quad (11)$$

Hyperparameter  $\lambda_{cons}$  is set to 1 following [20], whereas  $\lambda_p = 1e-6$  following [40] and  $\lambda_{ctx} = 1e-5$ .

## 4. Evaluation and Results

In this section, we demonstrate the effectiveness of SHIFT through a comprehensive evaluation of the target domain in

the absence of ground truth annotations. We provide extensive quantitative and qualitative results to highlight the strengths and limitations of our framework. Additionally, we conduct an ablation study to assess the effect of each loss, the sensitivity to the choice of the pseudo-label threshold, and the contributions of the individual modules to the overall framework.

**Datasets.** We use the following datasets in this work:

- **SURREAL** [42] is a large scale synthetic dataset with more than 6 million images of people with annotations for 25 joints. Generated from 3D sequences of human motion in an indoor setting, SURREAL features a diverse range of poses and viewpoints.
- **MINI-RGBD** [11] contains 12,000 synthetic infant images with annotations for 25 joints. Following standard training and evaluation settings, we train on splits ‘01’ to ‘10’ and validate on splits ‘11’ and ‘12’.
- **SyRIP** [14] contains 1,000 synthetic and 700 real images of infants with annotations for 17 joints. We train on the train split of 1,200 real and synthetic samples and evaluate on the test split of 500 real images.

#### 4.1. Implementation details

**Base Model Training.** We adopt the ResNet-101 [10] architecture as our backbone pose estimation model following the methodology in [47]. We first perform pre-training on the source dataset for 40 epochs, followed by an adaptation phase of 30 epochs in the target domain. The learning rate is initially fixed at  $1e-4$  with a multi-step decay by a factor of 0.1 after 5 epochs and 20 epochs. We use a batch size of 32 and the Adam optimizer [21] for all our experiments. To compare SHIFT with FiDIP [14], we retrain their models in a synthetic-to-real domain adaptation fashion, replacing the backbone with ResNet-101 [47]. The effect of varying the pseudo-label threshold on performance is thoroughly discussed in Section 4.4.

**Prior Training ( $\theta_p$ ).** We utilize the PoseNDF architecture [40] for training our prior model, following a multi-stage approach involving a mix of manifold and non-manifold poses. The number of non-manifold poses increases as the distance  $d$  of noisy poses from the plausible manifold grows. This allows the model to incorporate noisier poses by drawing more samples from the target-agnostic pose set as training progresses. We use a fixed batch size of 32 and train the model for 75, 100, and 150 epochs. The  $p_{enc}$  pose encoder consists of a 2 layer-MLP with a size of 6 for each orientation vector. We train the infant pose prior module in a supervised cross-dataset fashion on the constructed infant prior dataset. We adopt two training paradigms for the prior module: firstly, direct training on the target agnostic pose set; and secondly, initial training on the source dataset  $D_s$  followed by fine-tuning on our task’s agnostic pose set. We find that the former regimen outperforms the latter (more

details in supplementary).

**Kp2Seg Training ( $\mathcal{G}$ ).** We employ DC-GAN [34] as the backbone architecture. This module employs a linear layer followed by 5 convolutional layers to project the keypoints onto  $256 \times 256$  sized segmentation maps. The final output is upsampled to the desired size using bilinear interpolation. We keep a batch size of 64, a fixed learning rate of  $3e-4$ , and use the Adam optimizer for training the network. We train the network for a total of 200 epochs in a supervised manner. We extract segmentation masks from the synthetic images of the source (adult) domain using DeepLab-v3 [2] and maintain a binarization threshold of 0.5 across all evaluated cases. We use the SURREAL dataset for training this module due to the large number of samples present which makes it an ideal source dataset. *Existing infant datasets can’t be used for pre-training due to their very limited size and diversity in data.* We show results on different source datasets in Table 1 and in the supplementary.

Table 1. **Quantitative Results (PCK@0.05)** for SURREAL [42] → MINI-RGBD [11]. The best accuracies are highlighted in red and the second best accuracies are highlighted in blue.

Algorithm	SURREAL → MINI-RGBD							
	Head	Sld.	Elb.	Wrist	Hip	Knee	Ankle	Avg.
Source only	99.50	04.10	06.10	11.50	69.60	11.50	75.20	47.40
Oracle	100.00	99.70	97.40	75.00	92.60	86.10	84.30	89.20
RegDA [18]	90.80	15.10	24.50	26.90	73.80	24.60	62.10	37.80
UniFrame [20]	100.00	05.00	54.30	42.70	96.50	32.20	75.40	51.50
SHIFT	100.00	14.90	68.80	45.20	96.50	40.60	72.70	56.40

Table 2. **Quantitative Results (PCK@0.05)** for SURREAL [42] → SyRIP [14]. The best accuracies are highlighted in red and the second best accuracies are highlighted in blue.

Algorithm	SURREAL → SyRIP							
	Head	Sld.	Elb.	Wrist	Hip	Knee	Ankle	Avg.
Source only	52.40	35.60	23.50	27.10	32.90	14.20	24.70	26.30
Oracle	89.40	82.10	65.70	66.10	64.10	50.70	54.50	63.80
RegDA [18]	48.60	27.90	16.00	19.00	12.00	11.90	14.40	16.90
UniFrame [20]	54.40	47.50	13.50	31.10	50.60	26.00	36.50	34.20
SHIFT	53.40	46.10	34.20	38.70	51.10	31.20	37.60	39.80

Table 3. **Quantitative Results (PCK@0.05)** for SyRIP [14] → MINI-RGBD [11]. The best accuracies are highlighted in red and the second best accuracies are highlighted in blue.

Algorithm	Unsup	SyRIP → MINI-RGBD							
		Head	Sld.	Elb.	Wrist	Hip	Knee	Ankle	Avg.
Oracle	-	100.00	99.70	97.40	75.00	92.60	86.10	84.30	89.20
FiDIP [14]	x	24.80	54.10	88.30	83.60	19.50	88.40	74.60	68.10
SHIFT	✓	32.80	99.00	98.90	70.20	60.70	87.70	87.10	84.10

**Baselines and Metrics.** We evaluate the performance of our proposed method against the UDA-based frameworks RegDA [18] and UniFrame [20], as well as FiDIP [14] that performs supervised domain adaptation from synthetic to

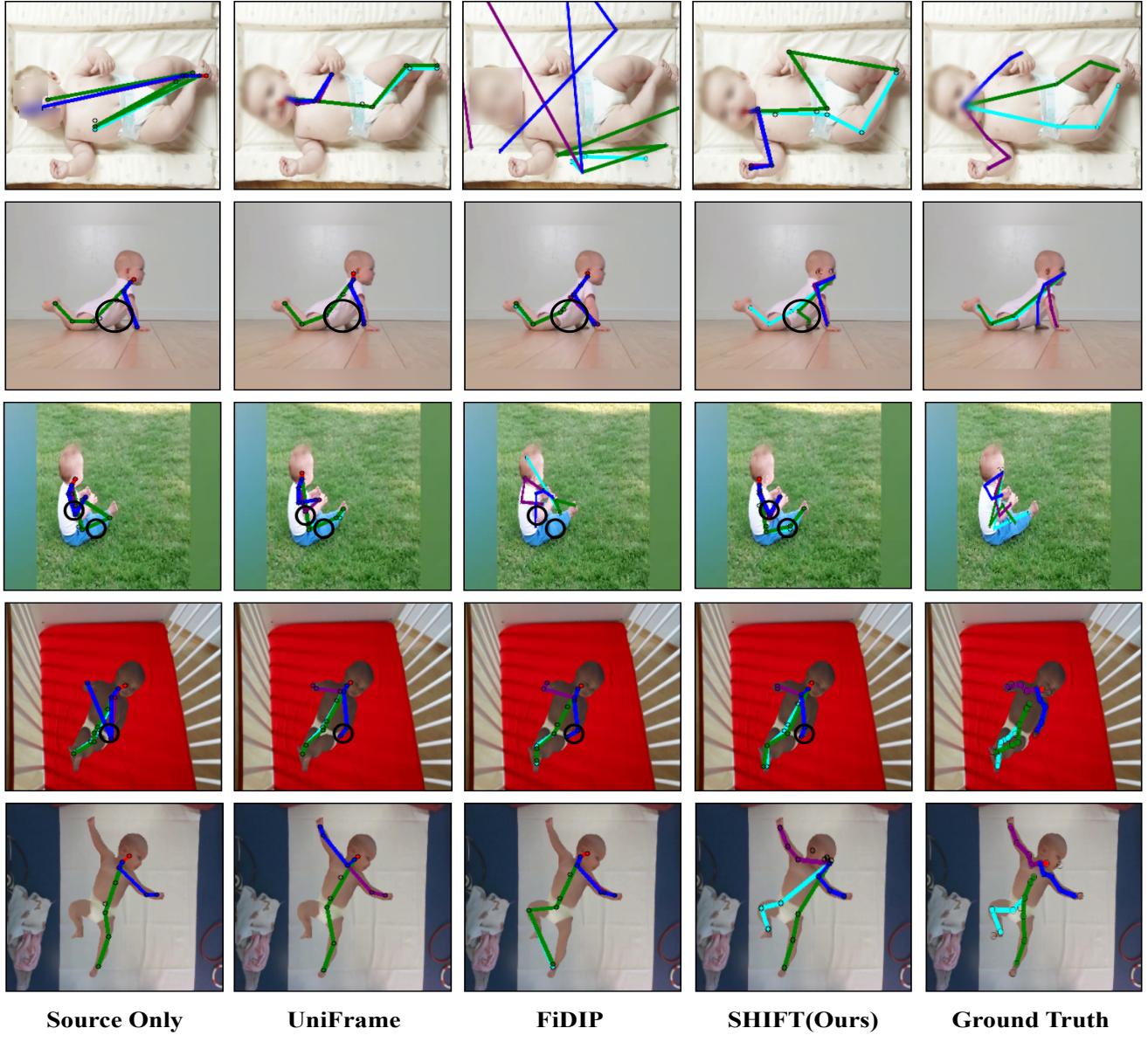


Figure 3. Qualitative results on **SURREAL** → **SyRIP** (top 3 rows) and **SURREAL** → **MINI-RGBD** (bottom 2 rows). From left to right: source only keypoints, keypoint predictions by UniFrame, predictions by FiDIP, predictions by SHIFT, and ground truth keypoints. As it can be seen above, the infant prior is essential to predict plausible poses in cases where other methods fail (top row). Further, our method can utilize context from visible regions to predict keypoints in self-occluded areas (2nd and 3rd row) while seamlessly adapting to different scenarios (4th and 5th row). ○ denotes the self-occluded regions in the images.

real infant images. To ensure a fair comparison, we retrain those models on the ResNet-101 architecture [47], which serves as the backbone in all cases. For a comprehensive baseline representation, we consider two additional baselines; *Oracle*, which is the upper bound obtained by training the model in a fully supervised manner on the target (infant) domain, and *Source-Only*, the lower bound resulting from direct inference of the unadapted source model on the target domain. Following prior works [18, 20], we use

the *Percentage of Correct Keypoints (PCK)* metric for all evaluations, which measures the percentage of keypoint detections within a specified distance from the true keypoints. All the accuracies reported henceforth use the **PCK@0.05** metric on 16 keypoints.

## 4.2. Quantitative Results

We evaluate SHIFT in two adult-to-infant UDA scenarios: SURREAL [42] → MINI-RGBD [11], and SUR-

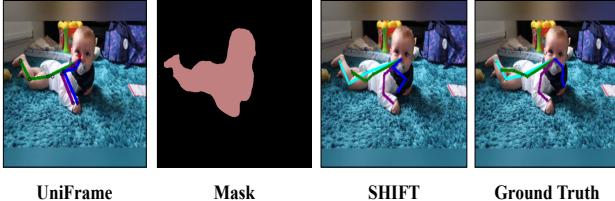


Figure 4. **Tackling Self-Occlusions: SURREAL → SyRIP.** UniFrame prediction (left panel) fails to correctly estimate significant portions of the lower back and left hand of the infant while SHIFT is able to reasonably do so. Ground truth (rightmost panel) and extracted mask (second from left panel) are also shown.

REAL [42] → SyRIP [14]; we also compare our method against state-of-the-art UDA methods [18, 20]. Additionally, we conduct an unsupervised evaluation for SyRIP → MINI-RGBD to compare against FiDIP [14], which incorporates a domain classifier to distinguish between real and synthetic images while fine-tuning on infant poses. Results are summarized in Table 1, Table 2 and Table 8, respectively. Among these methods, SHIFT achieves the highest performance across all cases, surpassing SOTA methods UniFrame [20] and RegDA [18] by approximately 5% and over 30%, respectively. Across the 16 individual keypoints of the infant, including shoulders (sld.), elbows (elb.), wrists, hips, knees, and ankles, SHIFT demonstrates superior performance in all categories except for the ankle in MINI-RGBD and the head and shoulders in SyRIP. Notably, despite FiDIP fine-tuning within the same dataset (SyRIP) in a fully supervised manner, our framework outperforms their approach without using any annotations.

These results underscore the limitations of solely learning discriminative features or performing prediction space alignment in scenarios with substantial domain gaps. SHIFT effectively bridges this gap by leveraging anatomical and contextual cues in the target domain through the integration of the infant pose prior and Kp2Seg modules, both of which can be directly trained offline. This comprehensive approach ensures robust adult-to-infant adaptation.

### 4.3. Qualitative Results

We also present some illustrative qualitative results from adapting SURREAL → MINI-RGBD and SURREAL → SyRIP in Figure 3. We further demonstrate how SHIFT handles occlusions effectively in Figure 4. In Figure 3, we illustrate SHIFT’s capability in capturing plausible infant poses while reasoning for occluded areas in various scenarios. Existing methods clearly lack anatomical understanding of infants in the top row while our method is able to reasonably predict keypoints while in the 2nd and 3rd row of Figure 3, UniFrame fails to correctly predict the keypoints around the infant’s elbows and wrists, whereas SHIFT can do so. Furthermore in Figure 4, SHIFT is able to reason-

ably estimate the lower back and left hand of the infant despite being heavily self-occluded. In all of these cases, it is evident that our framework SHIFT can seamlessly perform adult-to-infant domain adaptation in a data-efficient manner using no annotations.

### 4.4. Ablation Studies

**Effect of Loss Terms.** We perform a rigorous analysis to assess how each of the modules in our pipeline and each loss term in our adaptation objective affects the overall performance. Results in Table 4 demonstrate that the inclusion of the infant pose prior and the Kp2Seg module can lead to a notable increase in the overall performance. Further ablation results are present in the supplementary material.

Table 4. We analyse the effects of each loss term and module in this table for SURREAL [42] → MINI-RGBD [11].

Module	Loss Terms				PCK@0.05
	$\mathcal{L}_{\text{sup}}$	$\mathcal{L}_{\text{cons}}$	$\mathcal{L}_p$	$\mathcal{L}_{\text{ctx}}$	
Pre-Training	✓	✗	✗	✗	47.40
UDA [20]	✓	✓	✗	✗	51.50
UDA + Prior	✓	✓	✓	✗	53.60
SHIFT	✓	✓	✓	✓	<b>56.40</b>

**Effect of Pseudo-Label Threshold ( $\tau_c$ ).** Lastly, we study the effect that the pseudo-label threshold has on the framework. Results are listed in Table 5. It can be observed that both overly low or high values of the pseudo-label threshold, can negatively affect the adaptation process. This can be attributed to the fact that an overly low threshold degrades the quality of pseudo-labels while an overly high threshold filters out most pseudo-labels.

Table 5. We analyze the effects of different  $\tau_c$  values on performance (PCK@0.05). SURREAL [42] is the source dataset.

$\tau_c$	0.1	0.3	0.5	0.7	0.9
SyRIP [14]	35.00	39.00	<b>39.80</b>	37.50	35.10
Mini-RGBD [11]	53.00	53.70	<b>56.40</b>	54.10	53.50

## 5. Conclusion

We introduce SHIFT, an elegant framework for unsupervised pose estimation on infants. In contrast to existing analogous algorithms, SHIFT does not necessitate annotated training data, which is often cumbersome to obtain for infants. SHIFT utilizes the mean-teacher framework to provide effective self-supervision over the adaptation process with confident pseudo-labels, coupled with an infant manifold pose prior to act as an anatomical regularizer that enforces the student model to predict plausible poses, and a pose-image consistency module to provide additional contextual guidance to the model. Extensive experiments show

that our framework significantly outperforms existing state-of-the-art methods, thus providing superior performance on the challenging infant datasets.

## 6. Training Details and Adaptation Results

**Prior training paradigm.** We adopt two approaches for training our infant pose prior: the first approach includes training directly on the target agnostic dataset and the second approach includes training the prior on the source dataset and then fine-tuning (FT) on the target agnostic set. The results are as below:-

Table 6. Quantitative Results (PCK@0.05) for SHIFT against FiDIP [14].

Algorithm	SURREAL → MINI-RGBD							
	Head	Sld.	Elb.	Wrist	Hip	Knee	Ankle	Avg.
SHIFT w/o FT	96.00	29.20	48.90	34.40	86.10	43.50	75.00	52.80
SHIFT	100.00	14.90	68.80	45.20	96.50	40.60	72.70	56.40

Table 7. Quantitative Results (PCK@0.05) for SHIFT against FiDIP [14].

Algorithm	SURREAL → SyRIP							
	Head	Sld.	Elb.	Wrist	Hip	Knee	Ankle	Avg.
SHIFT w/o FT	43.40	40.20	35.20	38.40	49.20	29.20	36.80	38.10
SHIFT	45.60	45.00	35.90	38.00	51.40	31.40	32.00	39.00

Fine-tuning directly in a target agnostic setting provides better results than pre-training on source and fine-tuning on the target agnostic set. This suggests that our pre-training regimen is crucial towards preventing source knowledge forgetting; hence re-training the prior on the source dataset is not necessary.

**Synthetic Infant to Real Data Adaptation.** Using MINI-RGBD[11] as the source dataset results in unsatisfactory performance for both our method and the baseline. This is likely due to its limited diversity in infant poses and minimal inter-frame motion, which hinders effective pre-training for real images with high self-occlusion, as seen in SyRIP [14]. Despite SyRIP having fewer images, its diverse poses and scenarios make it a superior pre-training source.

Table 8. Quantitative Results (PCK@0.05) for SyRIP [14]→MINI-RGBD [11]. The best accuracies are highlighted in red and the second best accuracies are highlighted in blue.

Algorithm	Unsup	SyRIP → MINI-RGBD							
		Head	Sld.	Elb.	Wrist	Hip	Knee	Ankle	Avg.
Oracle	-	89.40	82.10	65.70	66.10	64.10	50.70	54.50	63.80
FiDIP [14]	X	52.20	21.30	22.40	14.40	33.20	26.00	23.90	27.55
SHIFT	✓	61.80	61.00	41.40	40.40	42.50	33.90	34.70	42.30

## 7. Additional Ablation Results

**Effect of Loss Terms.** We ablate each of the loss terms on the SyRIP [14] dataset. The strong role of Kp2Seg ( $\mathcal{G}(\cdot)$ ) is seen in dealing with self-occlusions.

Table 9. We analyse the effects of each loss term and module in this table for SURREAL [42] → SyRIP [14].

Module	Loss Terms				PCK@0.05
	$\mathcal{L}_{\text{sup}}$	$\mathcal{L}_{\text{cons}}$	$\mathcal{L}_p$	$\mathcal{L}_{\text{ctx}}$	
Pre-Training	✓	✗	✗	✗	26.30
UDA [20]	✓	✓	✗	✗	34.20
UDA + Prior	✓	✓	✓	✗	35.90
SHIFT	✓	✓	✓	✓	<b>39.80</b>

**Acknowledgments:** We gratefully acknowledge the support of NSF CMMI-2133084. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

## References

- [1] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7291–7299, 2017. 2
- [2] Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. Rethinking atrous convolution for semantic image segmentation, 2017. 2, 5, 6
- [3] Bowen Cheng, Bin Xiao, Jingdong Wang, Honghui Shi, Thomas S Huang, and Lei Zhang. Higherhrnet: Scale-aware representation learning for bottom-up human pose estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5386–5395, 2020. 2
- [4] Hao-Shu Fang, Shuqin Xie, Yu-Wing Tai, and Cewu Lu. Rmpe: Regional multi-person pose estimation. In *Proceedings of the IEEE international conference on computer vision*, pages 2334–2343, 2017. 2
- [5] Charles Fefferman, Sanjoy Mitter, and Hariharan Narayanan. Testing the manifold hypothesis, 2013. 2, 4
- [6] Riccardo Gatto and Sreenivasa Rao Jammalamadaka. The generalized von mises distribution. *Statistical Methodology*, 4(3):341–353, 2007. 5
- [7] Zigang Geng, Ke Sun, Bin Xiao, Zhaoxiang Zhang, and Jingdong Wang. Bottom-up human pose estimation via disentangled keypoint regression. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14676–14686, 2021. 2
- [8] Clément Godard, Oisin Mac Aodha, and Gabriel J Brostow. Unsupervised monocular depth estimation with left-right consistency. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 270–279, 2017. 3
- [9] Hanife Guney, Melek Aydin, Murat Taskiran, and Nihan Kahraman. A deep neural network based toddler tracking system. *Concurrency and Computation: Practice and Experience*, 34(14):e6636, 2022. 1

- [10] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 6
- [11] Nikolas Hesse, Christoph Bodensteiner, Michael Arens, Ulrich G Hofmann, Raphael Weinberger, and A Sebastian Schroeder. Computer vision for medical infant motion analysis: State of the art and rgb-d data set. In *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*, pages 0–0, 2018. 1, 2, 6, 7, 8, 9
- [12] Judy Hoffman, Eric Tzeng, Taesung Park, Jun-Yan Zhu, Phillip Isola, Kate Saenko, Alexei Efros, and Trevor Darrell. Cycada: Cycle-consistent adversarial domain adaptation. In *International conference on machine learning*, pages 1989–1998. Pmlr, 2018. 3
- [13] Xun Huang and Serge Belongie. Arbitrary style transfer in real-time with adaptive instance normalization, 2017. 3
- [14] Xiaofei Huang, Nihang Fu, Shuangjun Liu, and Sarah Ostadbash. Invariant representation learning for infant pose estimation with small data. In *2021 16th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2021)*, pages 1–8. IEEE, 2021. 1, 2, 6, 8, 9
- [15] Donald F Huelke. An overview of anatomical considerations of infants and children in the adult world of automobile safety design. In *Annual Proceedings/Association for the Advancement of Automotive Medicine*, page 93. Association for the Advancement of Automotive Medicine, 1998. 2
- [16] Eldar Insafutdinov, Leonid Pishchulin, Bjoern Andres, Mykhaylo Andriluka, and Bernt Schiele. Deepcut: A deeper, stronger, and faster multi-person pose estimation model. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part VI 14*, pages 34–50. Springer, 2016. 2
- [17] Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. Human3. 6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. *IEEE transactions on pattern analysis and machine intelligence*, 36(7):1325–1339, 2013. 1
- [18] Junguang Jiang, Yifei Ji, Ximei Wang, Yufeng Liu, Jianmin Wang, and Mingsheng Long. Regressive domain adaptation for unsupervised keypoint detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6780–6789, 2021. 1, 3, 6, 7, 8
- [19] Alex Kendall, Matthew Grimes, and Roberto Cipolla. Posenet: A convolutional network for real-time 6-dof camera relocalization. In *Proceedings of the IEEE international conference on computer vision*, pages 2938–2946, 2015. 2
- [20] Donghyun Kim, Kaihong Wang, Kate Saenko, Margrit Betke, and Stan Sclaroff. A unified framework for domain adaptive pose estimation. In *European Conference on Computer Vision*, pages 603–620. Springer, 2022. 1, 2, 3, 5, 6, 7, 8, 9
- [21] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 6
- [22] Jürgen Konczak and Johannes Dichgans. The development toward stereotypic arm kinematics during reaching in the first 3 years of life. *Experimental brain research*, 117:346–354, 1997. 2
- [23] Lingdong Kong, Youquan Liu, Lai Xing Ng, Benoit R Cottereau, and Wei Tsang Ooi. Openess: Event-based semantic scene understanding with open vocabularies. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15686–15698, 2024. 5
- [24] Jogendra Nath Kundu, Phani Krishna Uppala, Anuj Pahuja, and R Venkatesh Babu. Adadepth: Unsupervised content congruent adaptation for depth estimation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2656–2665, 2018. 3
- [25] Julianne LaChance, William Thong, Shruti Nagpal, and Alice Xiang. A case study in fairness evaluation: Current limitations and challenges for human pose estimation. In *Association for the Advancement of Artificial Intelligence 2023 Workshop on Representation Learning for Responsible Human-centric AI (R2HCAI), Washington, DC*, 2023. 1
- [26] Chen Li and Gim Hee Lee. From synthetic to real: Unsupervised domain adaptation for animal pose estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1482–1491, 2021. 3
- [27] Cheng Li, A. Pourtaherian, L. van Onzenoort, W. E. Tjon a Ten, and P. H. N. de With. Infant facial expression analysis: Towards a real-time video monitoring system using r-cnn and hmm. *IEEE Journal of Biomedical and Health Informatics*, 25(5):1429–1440, 2021. 1
- [28] Caio Mucchiani, Zhichao Liu, Ipsita Sahin, Jared Dube, Linh Vu, Elena Kokkoni, and Konstantinos Karydis. Closed-loop position control of a pediatric soft robotic wearable device for upper extremity assistance. In *31st IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*, pages 1514–1519, 2022. 1
- [29] Caio Mucchiani, Zhichao Liu, Ipsita Sahin, Elena Kokkoni, and Konstantinos Karydis. Robust generalized proportional integral control for trajectory tracking of soft actuators in a pediatric wearable assistive device. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2023. 1
- [30] Alejandro Newell, Kaiyu Yang, and Jia Deng. Stacked hourglass networks for human pose estimation, 2016. 2
- [31] Alejandro Newell, Zhiao Huang, and Jia Deng. Associative embedding: End-to-end learning for joint detection and grouping. *Advances in neural information processing systems*, 30, 2017. 2
- [32] Qucheng Peng, Ce Zheng, and Chen Chen. Source-free domain adaptive human pose estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4826–4836, 2023. 3
- [33] Leonid Pishchulin, Eldar Insafutdinov, Siyu Tang, Bjoern Andres, Mykhaylo Andriluka, Peter V Gehler, and Bernt Schiele. Deepcut: Joint subset partition and labeling for multi person pose estimation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4929–4937, 2016. 2
- [34] Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks, 2016. 5, 6

- [35] Dimitrios Sakkos, Kevin D Mccay, Claire Marcroft, Nicholas D Embleton, Samiran Chattopadhyay, and Edmond SL Ho. Identification of abnormal movements in infants: A deep neural network for body part-based prediction of cerebral palsy. *IEEE Access*, 9:94281–94292, 2021. 1
- [36] Giuseppa Sciortino, Giovanni Maria Farinella, Sebastiano Battiato, Marco Leo, and Cosimo Distante. On the estimation of children’s poses. In *Image Analysis and Processing-ICIAP 2017: 19th International Conference, Catania, Italy, September 11-15, Proceedings, Part II 19*, pages 410–421. Springer, 2017. 2
- [37] Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang. Deep high-resolution representation learning for human pose estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5693–5703, 2019. 2
- [38] D Sutherland. The development of mature gait. *Gait & posture*, 6(2):163–170, 1997. 2
- [39] Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. *Advances in neural information processing systems*, 30, 2017. 2, 3, 4
- [40] Garvita Tiwari, Dimitrije Antić, Jan Eric Lenssen, Nikolaos Sarafianos, Tony Tung, and Gerard Pons-Moll. Pose-ndf: Modeling human pose manifolds with neural distance fields. In *European Conference on Computer Vision*, pages 572–589. Springer, 2022. 2, 4, 5, 6
- [41] Jonathan J Tompson, Arjun Jain, Yann LeCun, and Christoph Bregler. Joint training of a convolutional network and a graphical model for human pose estimation. *Advances in neural information processing systems*, 27, 2014. 3
- [42] Gul Varol, Javier Romero, Xavier Martin, Naureen Mahmood, Michael J Black, Ivan Laptev, and Cordelia Schmid. Learning from synthetic humans. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 109–117, 2017. 1, 2, 6, 7, 8, 9
- [43] Michael Wan, Shaotong Zhu, Lingfei Luan, Gulati Prateek, Xiaofei Huang, Rebecca Schwartz-Mette, Marie Hayes, Emily Zimmerman, and Sarah Ostadabbas. Infanface: Bridging the infant-adult domain gap in facial landmark estimation in the wild. In *2022 26th International Conference on Pattern Recognition (ICPR)*, pages 4486–4492. IEEE, 2022. 1
- [44] Michael Wan, Xiaofei Huang, Bethany Tunik, and Sarah Ostadabbas. Automatic assessment of infant face and upper-body symmetry as early signs of torticollis. In *IEEE 17th International Conference on Automatic Face and Gesture Recognition (FG)*, pages 1–6. IEEE, 2023. 1
- [45] Shih-En Wei, Varun Ramakrishna, Takeo Kanade, and Yaser Sheikh. Convolutional pose machines. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 4724–4732, 2016. 2
- [46] Yuan Wu, Diana Inkpen, and Ahmed El-Roby. Dual mixup regularized learning for adversarial domain adaptation. In *Computer Vision-ECCV 2020*, pages 540–555. Springer, 2020. 3
- [47] Bin Xiao, Haiping Wu, and Yichen Wei. Simple baselines for human pose estimation and tracking. In *Proceedings of the European conference on computer vision (ECCV)*, pages 466–481, 2018. 2, 6, 7
- [48] Tete Xiao, Yingcheng Liu, Bolei Zhou, Yuning Jiang, and Jian Sun. Unified perceptual parsing for scene understanding. In *Proceedings of the European conference on computer vision (ECCV)*, pages 418–434, 2018. 5
- [49] Yufei Xu, Jing Zhang, Qiming Zhang, and Dacheng Tao. Vitpose: Simple vision transformer baselines for human pose estimation. *Advances in neural information processing systems*, 35:38571–38584, 2022. 2
- [50] Tinghui Zhou, Matthew Brown, Noah Snavely, and David G Lowe. Unsupervised learning of depth and ego-motion from video. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1851–1858, 2017. 3
- [51] Zhuoran Zhou, Zhongyu Jiang, Wenhao Chai, Cheng-Yen Yang, Lei Li, and Jenq-Neng Hwang. Efficient domain adaptation via generative prior for 3d infant pose estimation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 41–49, 2024. 1, 2