SHIFT: LEVERAGING SYNTHETIC ADULT DATASETS FOR INFANT POSE ESTIMATION

By

Hannah Loi Dela Cruz-Yuan

A capstone project submitted for Graduation with University Honors

May 7, 2026

University Honors

University of California, Riverside

APPROVED

Dr. Amit K. Roy-Chowdhury

Department of Electrical and Computer Engineering

Dr. Begoña Echeverria, Howard H Hays, Jr. Chair

University Honors

Abstract

Cerebral palsy is a neurological disorder characterized by abnormal, spontaneous limb movements in infants. Early detection is crucial for timely treatment, but such assessments are challenging and rigorous, requiring extensive training to objectively identify abnormal movement patterns. Pose estimation models enable automated, markerless joint tracking, which can be used to detect and quantify movements patterns. However, most models are built using adult subjects, and those for infants require labeled training data, which is very laborious to generate and difficult to access. To address these challenges, we built an adaptive method called SHIFT: Leveragine **S**ynt**H**etic Adult Datasets for **I**n**F**an**T** Pose Estimation. SHIFT incorporates innovative techniques like the mean teacher framework, an infant manifold pose prior, and a novel visibility consistency module to effectively adapt a pre-trained 2D adult pose estimation model to infants. We performed extensive experiments on multiple pose estimation benchmarks, including unsupervised domain adaptation and supervised methods. We demonstrate that SHIFT outperforms unsupervised domain adaptation methods by an increase of 5% (51% vs 56%, respectively) and supervised methods by an increase of 16% (68% vs 84%, respectively) in prediction accuracy. Building on this foundation, we conduct additional ablation studies to assess the optimal weighting of the manifold pose prior and keypoint segmentation modules, ensuring that anatomical constraints enhance rather than distort predictions during the adaptation process. We now propose SHIFT as an infant pose estimation method independent of labeled target domain data, which we hope will enable efficient assessment of disorders such as cerebral palsy using computer vision.

Acknowledgments

This research was supported in part by the National Science Foundation (NSF) under Grant No. CMMI-2133084. Hannah Dela Cruz-Yuan was supported as a trainee by the National Institutes of Health (NIH) under Grant No. 5T34GM149470-02. This work was conducted as an expansion of the first author's University Honors project at UC Riverside and builds upon research previously published at the 8th ABAW Workshop at CVPR 2025 [1], on which she served as a co-first author.

The authors would like to thank Sarosij Bose, Arindam Dutta, and Elena Kokkoni for their collaborative contributions to the original framework. We also extend our gratitude to Dr. Amit K. Roy-Chowdhury, director of the UCR Vision and Learning Lab for his invaluable guidance and support throughout the duration of this project.

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 or the National Institutes of Health.

Table of Contents

1 Introduction

The mean teacher framework is a surprisingly effective mechanism for semi-supervised machine learning. We will begin by reviewing its methodology. First introduced by Tarvainen and Valpola in 2017 [23], this method utilizes a dual-model training architecture which involves two identical neural networks, the "student" and "teacher" models. At the core of this method, while the student model learns to correctly analyze the training data via traditional backpropagation machine learning methods, the teacher model does not learn, but rather represents a moving reflection of the student model's training history [23]. Uniquely, the mean teacher framework uses unlabeled training data, and depends on the stability of the teacher model to instruct the predictions of the student model [23, 13]. This semi-supervised mechanism will be further elaborated later.

This work utilizes a variation of the mean teacher framework which emphasizes transfer learning, and more specifically, domain adaptation. We will define these terminology and their implications for this work. By transfer learning, we refer to the process of pre-training both models on a large-scale dataset to initialize the models with stable parameters (weights) that can be further fine-tuned [26, 11]. We are interested in the subset of transfer learning that is domain adaptation (DA), which refers to the process of training a model on one dataset (or source domain) and then adapting that model to a new dataset (or target domain) [12, 13]. In the context of the mean teacher framework, after pretraining both models on the large-scale dataset from a "source" domain, the pair of models is then exposed to the "target" domain and trained in the manner of the mean teacher methodology [13]. Of additional benefit, using this framework eliminates the need for curating ground-truth labels on the target domain dataset [23, 12]. We will further elaborate on the benefit of this characteristic shortly.

2 Related Works

2.1 The Mean-Teacher Paradigm in Computer Vision

Having briefly covered the introduction and framework of the mean teacher network, we will now proceed to discuss relevant applications. In the original paper, the authors evaluated the performance of the mean teacher framework on the datasets CIFAR-10 [14] and ImageNet2012

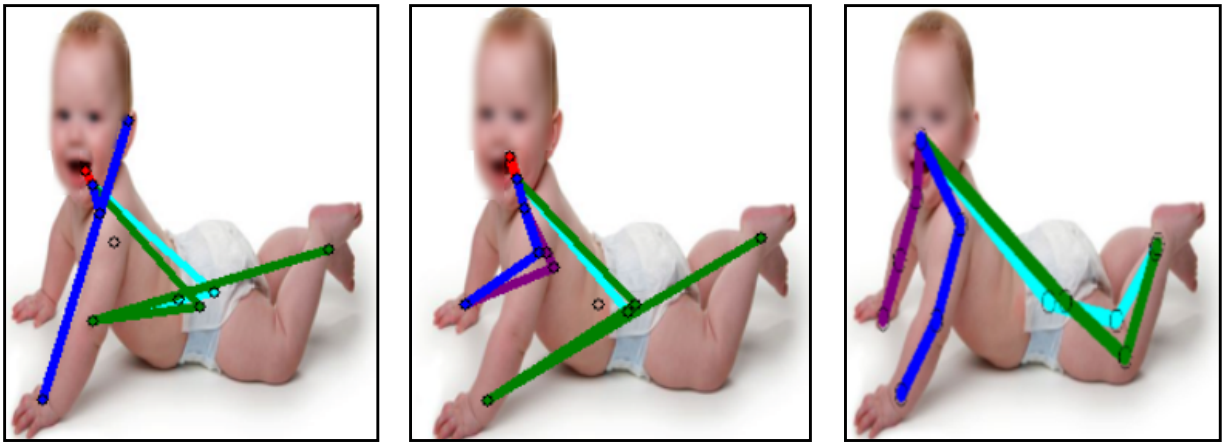

Figure 1: **Need for unsupervised domain adaptive infant pose estimation.** From left to right: keypoint predictions from a baseline adult human pose estimation model, predictions from a SOTA UDA model (UniFrame) [13], and predictions from our method, `SHIFT`. Direct application of adult models fails due to the domain shift; `SHIFT` effectively adapts by accounting for the self-occluded pose distribution of infants.

[19], which are composed of images and their corresponding class labels. Their work demonstrates that this methodology is able to effectively learn relevant information from the images in order to correctly assign a class label [23].

2.2 Domain Adaptive Pose Estimation

Building on their work, Kim et al. in 2022 [13] extends the utility of the mean teacher framework to domain adaptive pose estimation. Pose estimation is an extension of image classification which moves beyond assigning labels to a whole image to assigning labels to objects in the image, or more specifically, identifying and assigning labels to the joints of the subject [16, 22]. Leveraging the semi-supervised nature of the mean teacher framework, the authors successfully achieve "state-of-the-art performance under various domain shifts" [13]. The source-target domain pairs which were evaluated include datasets of humans, hands, dogs, and sheep. Furthermore, they demonstrate the ability "to mitigate domain shift on diverse tasks and even unseen domains and objects (e.g., trained on horse and tested on dog)" [13]. Their work critically demonstrates the effectiveness of this methodology on image analysis tasks beyond classification.

Of the approaches discussed thus far, again note that access to a labeled source domain dataset

is key to achieving optimal accuracy on the unlabeled target domain dataset [12, 13]. However, obtaining high-quality ground truth joint annotations for pose estimation training can be painstakingly laborious and time consuming. Furthermore, ethical privacy issues regarding the human pose image datasets necessitate measures to limit the distribution of and access to these datasets [15]. In response to these valid limitations, Raychaudhuri et al. in 2023 [18] present an additional refinement to the application of the mean teacher framework for domain adaptation in human pose estimation. Specifically, the authors facilitate adaptation to the target domain dataset by incorporating the use of a human pose prior in place of the labeled source domain dataset [18]. This prior model has been previously trained to capture the range of poses and motions which humans assume, so as to regulate and refine the predictions of the model in training [24, 18]. Their methodology "achieves comparable performance to recent state-of-the-art methods that use source data for adaptation" [18]. Their work distinctly pivots away from dependence on labeled datasets, expanding the use of this methodology for applications where privacy concerns and limited access to training data would previously limit research efforts.

Having discussed the development of the mean teacher methodology for domain adaptation, specifically in the context of pose estimation, and having introduced the use of a pose prior for source-free domain adaptation, we now turn to the application of these tools towards a new domain adaptation task. In this work, we seek to extend the applications of the mean teacher framework one step further. In addition to adapting the model to a new human pose dataset, we will also adapt to a new life stage: **the human infant**. By infant, we refer to persons under the age of two.

2.3 Challenges in the Infant Human Domain

The human infant domain presents multiple challenges for pose estimation tasks, as demonstrated in Figure 1. Primarily, the range of infant poses occupies a distinctly separate distribution from adult humans [10]. Adult humans typically engage in upright postures and actions: walking, running, sitting, biking, etc. On the contrary, the infant human spends the majority of time in a horizontal position: lying down, rolling, crawling, etc [21]. The majority of these poses contain multiple self-occlusions, where the infant's limbs may interact and cross over other parts of the

body in such a way as to cause severe prediction difficulties for a pose estimation model [8].

From an ethical point of view, the infant domain presents additional challenges due to the privacy concerns regarding the photo capture of individuals without their consent (although guardians may grant approval), and the lack of monitoring of the distribution and use of such datasets [15].

However, tools for infant pose estimation may hold great potential. In the healthcare domain, studious monitoring of the behavior and mobility of infants has tremendous benefits for early detection of neurological challenges which may be treatable and properly alleviated, as in the case of cerebral palsy [20, 6]. Other applications of infant pose estimation include safety and wellness practices, such as at-home monitoring to reinforce infant safety in situations where guardian supervision is limited [5].

Building on the existing knowledge base previously elaborated, we propose a methodology for infant pose estimation which leverages the strengths of the mean teacher framework, domain adaptation for pose estimation, and a pose prior for training reinforcement to collectively achieve comparable performance on unseen infant images without necessitating access to and use of privacy-protected manually curated and labeled infant image datasets [9, 18].

3 Methodology

3.1 Overview of the Mean-Teacher Framework

We now propose `SHIFT`, a novel semi-supervised domain adaptation framework for adapting, or `SHIFT`ing, a pose estimation model from 2D adult human poses to 2D infant human poses. As previewed, we will now elaborate on the semi-supervised learning approach which `SHIFT` utilizes, as illustrated in Figure 2. In keeping with the principles of transfer learning for domain adaptation, we begin by training the model on a labeled dataset of adult humans from the source domain [26, 11]. After this "pre-training" phase is complete, we move to the mean teacher phase [23].

To set up the mean teacher framework, we initialize two copies of the pre-trained model, designated as the teacher and student models. Each training iteration, we present the teacher and the student models with a pair of closely related images to learn. The teacher model receives the original copy of the image, while the student model receives an altered version of the image that has

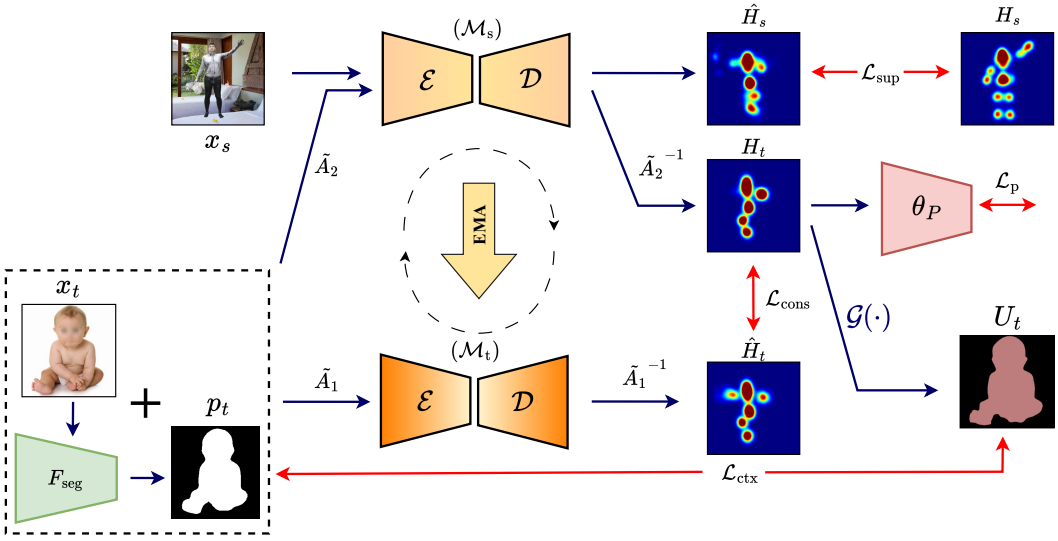

Figure 2: **Framework Overview.** SHIFT utilizes the Mean-Teacher framework to update the teacher model via EMA. It incorporates an infant pose prior for anatomical plausibility and a keypoint segmentation mapping module that aligns predicted poses with pseudo-segmentation masks to handle self-occlusions.

undergone shearing, rotation, and other warping transformations [13]. Each model returns its predictions of the joint annotations in the two images. After reverse transforming the student model predictions to the original coordinate system, the predictions are compared to one another and the error, or loss, is calculated [23]. With each new training iteration, the models work to minimize the error calculated between their predictions. Herein lies the crux of the algorithm, as the dual-model system self-learns robustness to distortions and implicitly learns the pose distribution of the target domain dataset [12, 13].

3.2 The Anatomical Pose Prior.

However, the demands of this adaptation task require further reinforcement to effectively overcome the domain gap between adult human and infant human pose distributions. We mitigate this by supplementing the iterative semi-supervised training architecture with a human pose prior, incorporating the tools proposed by Raychaudhuri et al. (2023) [18]. This human pose prior serves as an anatomical guide which effectively constrains the model to learning an anatomically plausible distribution of poses which the infant subject may assume [24, 18]. Critically, this pose prior model is trained on a human dataset auxiliary to the one which the model is learning, so as to avoid

contamination in the training process, which could lead to data leakage and overfitting.

3.3 The Context-Aware Segmentation Module.

We introduce a novel module to the SHIFT training paradigm, which has not previously been discussed. To address the aforementioned difficulty of self-occlusion prevalent in the infant human pose distribution, we incorporate a segmentation model which effectively delineates the boundaries of the infant subject in the image [3]. This border control serves to further restrict the predictions of the model to the area within the silhouette of the infant.

3.4 The SHIFT Training Architecture.

Having introduced the conceptual framework of the SHIFT methodology, we will proceed to briefly elaborate on the mechanistic architecture, as implemented for a 2D pose estimation model.

To prepare for domain adaptation to infant humans, we pre-train a model f_s on the source domain, an adult human dataset [26]. This pretraining occurs in a supervised manner, such that f_s accesses both the images and the joint annotations for each image in this dataset during training. The pretraining protocol utilizes a simple and classic mean squared error loss between the prediction heatmaps and the ground truth heatmaps, which we denote as L_{sup}, to back propagate f_s during this stage of the training process [25]. After a preset number of supervised pretraining epochs, f_s transitions to the semi-supervised self-training stage via the mean teacher methodology [23].

Upon completion of the pretraining stage, a teacher model f_{tea} and a student model f_{stu} are instantiated and loaded with the f_s weights, which are finetuned to the source domain adult human pose distribution. As mentioned previously, the weights of the teacher and student models are updated each epoch by two distinctly separate mechanisms. The parameters student model are updated by evaluating the outputs of the teacher and student models, calculating a loss function output, and backpropagating those results. The parameters of the teacher model are updated as an exponential moving average of the parameters of the student model, such that the teacher represents a moving history of the student's training progress [23]. This mechanism is formulated as follows:

$$\bar{\theta}_t = \alpha \bar{\theta}_{t-1} + (1-\alpha)\theta_t \tag{1}$$

where α represents the smoothing coefficient.

Setting α to 0.999 stabilizes the teacher model by limiting updates to only 0.1% per epoch, ensuring that the general rules of human anatomy learned in the pretraining stage are not lost during the semi-supervised training stage [13]. Otherwise, the sudden domain SHIFT from labeled adult humans to unlabeled infant humans may be a challenging threshold to cross and can cause the model training to become unstable without a high α value.

We generate two different views of incoming target images x_t by performing augmentations A_1 and A_2 on the inputs of the student model f_{stu} and the teacher model f_{tea}, respectively. These augmentations, which include random rotations, color changes, blurring, shearing, and translating of the infant human images, serve to enforce consistency regularization [13]. By distorting the inputs of the teacher and student models, the models learn to overcome the image distortions and prioritize anatomical consistency.

Similar to [13], we further tamper the input image to the student model by selectively patching out keypoints for which the teacher model f_{tea} produces the highest activations, by utilizing a patching operation

$$\hat{x}_t = P(\bar{A}_2(x_t)) \tag{2}$$

By identifying and randomly obscuring high-visibility joints which the teacher model detected, we force the student model to refine its predictions for occluded keypoints. In this manner, we synthesize self-occlusions and train the model to become robust to self-occlusions naturally occurring in the infant pose distribution.

Keeping with the semi-supervised nature of the mean teacher architecture, we take the high-confidence outputs of the teacher model and treat them as pseudo-labels with which to evaluate the outputs of the student model. In other words, eliminating the use of manually curated labels, we exchange those for the outputs of the teacher model, which has been stabilized with training on the adult human dataset [12, 13]. We follow a standard mean squared error function to generate this loss L_{cons} in order to back propagate the student model and effectively enforce consistency

between the student model's predictions and the teacher model's pseudo-labels [23]. We use the ResNet-101 architecture [7], as per the methodology in [12]. Pretraining occurs in a supervised fashion on the adult dataset for 40 epochs, followed by mean teacher adaptation to the target infant dataset for a subsequent 30 epochs.

3.5 Training the Anatomical Pose Prior.

Adjacent to the mean teacher architecture, as previously mentioned we incorporate a human pose prior to account for the anatomically distinct pose distributions of adult humans and infant humans. This prior models all physically plausible poses as a zero-level set [24], and penalizes anatomically implausible poses, forcing the student model to adjust its outputs to reflect the learned constraints of human anatomy [18]. Employing a cross-dataset training approach, the prior is trained on an auxiliary dataset , which is distinctly different from the dataset which it will evaluate during the mean teacher training phase. This separation of datasets effectively prevents data leakage.

To effectively train the pose prior for distinguishing between anatomically plausible and implausible poses, we take the 2D pose vectors from this auxiliary pose dataset and convert them into 2D orientation vectors to more accurately capture the spatial relationships between the joints [24]. Taking these anatomically plausible poses as the ground truth basis, the zero-level set, we generate noisy poses by sampling noise from the von Mises distribution [4]. We effectively compose a pose dataset composed of orientation vectors and a label binarization into anatomically plausible and implausible categories.

Following [18], we train the pose prior with a multi-stage approach marked by differences in the pose distribution of the synthetic training samples. In the first stage of training, the model is presented with a set of plausible poses and a set of distinctly implausible poses. With each new stage in training, the second set gradually transitions into less distinctly implausible poses. Through this approach, the model successfully learns to initially identify plausible poses and then gradually fine tunes to the distinctive details of those plausible poses which differentiate it from subtly implausible poses.

During mean teacher adaptation, we leverage this trained prior to output an anatomical plausi-

bility score of the student model output, based on the distance between the predicted pose and the learned plausible pose manifold [24]. To incorporate this score into the backpropagation update of the student model, we denote the pose loss function as:

$$\mathscr{L}_p = \frac{1}{N_t'} \sum_{x_t \in D_t} \theta_p(T(f_{stu}(x_t))) \tag{3}$$

Here, 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 [18]. The incorporation of this pose prior module effectively guides the student model towards a richer understanding of the distinct anatomy of the infant human. We utilize the PoseNDF architecture [24] and train the model for 75, 100, and finally 150 epochs by each stage.

3.6 Training the Segmentation Module.

The third and final module of the mean teacher training framework highlights context aware adaptation, such that the student model learns to utilize spatial information to more accurately align its predictions to what is anatomically plausible. As discussed previously, the pose distribution of the infant human contains frequent occurrences of self-occlusions due to the horizontal positions infants often assume. To further overcome the challenges of self occlusion in addition to domain adaptation, we utilize segmentation masks to provide additional contextual guidance to the student model during training. In other words, we train the student model to further restrict its joint predictions to the space within the boundaries of the infant in the image.

Specifically, we employ DeepLabv3 [2] to extract foreground-background segmentation masks for each of the target training images. We then use the pseudo masks as a ground truth to compare the student model predictions against. We transform the student predictions from a vector of keypoints into a segmentation mask. This is achieved by training a segmentation encoder module in a supervised fashion on a synthetic auxiliary dataset composed of poses and corresponding segmentation masks [3]. To incorporate this score into the backpropagation update of the student model, we denote the segmentation mask loss function as:

$$\mathscr{L}_{seg} = \frac{1}{N_s} \sum_{i=1}^{N_s} \sum_{j=1}^{H'' \times W''} -p_{s,j}^{i} \log(u_{s,j}^{i}) \tag{4}$$

where N_s represents the batch size of source images. Critically, this module is domain agnostic, meaning that it does not require ground truth segmentation masks during the mean teacher adaptation process.

To train the segmentation module, we utilize the DC-GAN architecture [17]. We use an adult dataset [26] which contains ground truth keypoint labels and segmentation masks of the human subject to train the module to effectively upsample a sparse keypoint heatmap to a dense body segmentation map.

3.7 The Weighted Adaptation Objective.

Combining these modules, the SHIFT training architecture consists of a pretraining phase on the source domain dataset, followed by a multi-module mean teacher training phase coupled with anatomical regularization enforced by the pose prior module and the keypoint segmentation module. Our student model f_{stu} is thus trained using the weighted adaptation objective:

$$\mathscr{L}_{total} = \mathscr{L}_{sup} + \lambda_{cons}\mathscr{L}_{cons} + \lambda_{prior}\mathscr{L}_{prior} + \lambda_{seg}\mathscr{L}_{seg} \tag{5}$$

We will now proceed to detail the specific implementation of the SHIFT domain adaptation framework in the context of the adult human and infant human datasets utilized in this work. We provide extensive quantitative and qualitative results to highlight the strengths and weaknesses of our label-free semi-supervised learning architecture. Additionally, we conduct an ablation study to assess the effect of each training loss function, the performance sensitivity to the pseudo-label threshold value, and the contribution of each individual module to the overall framework.

4 Experiments and Results

4.1 Experimental Setup

Datasets: We use the following datasets for our experiments.

SURREAL [26] is a large-scale dataset of single-person images of adult humans generated

Table 1: **Quantitative Results (PCK@0.05)** for **SURREAL** [26] → **MINI-RGBD** [8]. 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 [12]	90.80	15.10	24.50	26.90	73.80	24.60	62.10	37.80
UniFrame [13]	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** [26] → **SyRIP** [9]. 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 [12]	48.60	27.90	16.00	19.00	12.00	11.90	14.40	16.90
UniFrame [13]	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

from 3D sequences of human motion sequences, which have been overlaid with a synthetic skin and placed on randomly curated indoor backgrounds. Composed of more than six million images and annotations for 25 joints per image, SURREAL features a diverse range of poses and viewpoints.

Mini-RGBD [8] is a small dataset of single-person images of infant humans generated from 3D sequences of infant motion sequences, which have been overlaid with a synthetic skin and placed on a synthetic infant cot background. The dataset is split into twelve sub-sections, each consisting of one thousand images from a distinct video capture sequence. As all the infants are lying in a supine position, the poses from this dataset are more limited. Composed of twelve thousand images and annotations for 25 joints per image, Mini-RGBD features a relatively sizable dataset with limited poses. Following standard training and evaluation settings, we train on splits '01' to '10' and evaluate on splits '11' and '12'. The authors note that splits 11 and 12 are the most

Table 3: **Quantitative Results (PCK@0.05)** for **SyRIP** [9]→ **MINI-RGBD** [8]. 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 [9]	✗	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

difficult sub-sections due to the high frequency of self-occlusion.

SyRIP [9] is a significantly smaller dataset of single-person images, of real and synthesized infant humans. Specifically, seven hundred images are curated from online archives (videos, web search images, etc) and one thousand images are synthesized by augmenting the pose distribution of the real infants and then overlaying a synthetic skin and placing the synthetic infant on randomly curated indoor backgrounds, similar to SURREAL. We train on the 1000 synthetic images and 200 real infants, and reserve 500 real infant images for evaluation.

Baselines and Bounds: To accurately evaluate the performance of SHIFT, we prepare and train a selection of existing adult and infant pose estimation models as baselines and bounds.

To compare the improvements of SHIFT in domain adaptation, we evaluate against two key benchmarks, RegDA [12] and UniFrame [13]. RegDA (regressive domain adaptation) employs an adversarial regression strategy to identify and stabilize disparities across the source and target domains. UniFrame represents the baseline mean teacher adaptation, without the additional support modules utilized by SHIFT.

To compare the performance of SHIFT as an infant pose estimation model, we evaluate against an existing infant pose estimation model, FiDIP [9]. The FiDIP architecture consists of a supervised learning mechanism combined with an adversarial training approach to effectively learn the distinguishing features of infants across both synthetic and real datasets. For comparability to SHIFT, we re-train the FiDIP model using a synthetic-to-real domain adaptation approach, replacing the backbone with ResNet-101 [7, 27].

Finally, we maintain an upper bound and a lower bound by preparing two additional models

which take from the SHIFT architecture but are restricted to their respective datasets. We prepare an oracle model as the upper bound by training the model in a fully supervised manner on the target infant domain. We prepare a source-only model as the lower bound by training the model in a fully supervised manner on the source adult domain.

Evaluation Metric: Following prior works [12, 13, 18], we use the Percentage of Correct Keypoints (PCK) metric to evaluate all models. Therefore, all accuracies reported in the following tables use the PCK@0.05 metric on sixteen bodypart keypoints.

To make full use of the only publicly available infant datasets, we evaluate SHIFT adaptation for two scenarios: from the SURREAL [26] adult dataset to the MiniRGBD [8] infant dataset and from the SURREAL adult dataset to the SyRIP [9] infant dataset.

4.2 Quantitative Analysis

SURREAL $\rightarrow$ Mini-RGBD: In Table 1 we evaluate the SHIFT adaptation from SURREAL to MiniRGBD. SHIFT achieves the highest performance across all cases, surpassing SOTA methods RegDA and UniFrame by 30% and 5%, respectively. Across the 16 individual keypoints of the infant body which are evaluated, SHIFT demonstrates superior performance in nearly all categories, except for the ankle joint. These performance improvements are qualitatively illustrated in Figure 3, where SHIFT accurately captures plausible infant poses and effectively deduces anatomical rationale for occluded areas, in a variety of scenarios.

SURREAL $\rightarrow$ SyRIP: In Table 2 we evaluate the SHIFT adaptation from SURREAL to SYRIP. SHIFT surpasses RegDA and UniFrame by 20% and 5%, respectively. Across the 16 individual keypoints of the infant body which are evaluated, SHIFT demonstrates superior performance in nearly all categories, except for the head and shoulder joints. These performance improvements are qualitatively illustrated in Figure 3. The other methods lack an anatomical context for the infant pose distribution, whereas SHIFT effectively resolves self-occlusions and outputs plausible and mostly accurate joint keypoint predictions.

Infant-to-Infant Adaptation: We separately compare the performance of the SHIFT and FiDIP architectures when trained in a domain adaptation fashion from the SyRIP infant dataset to

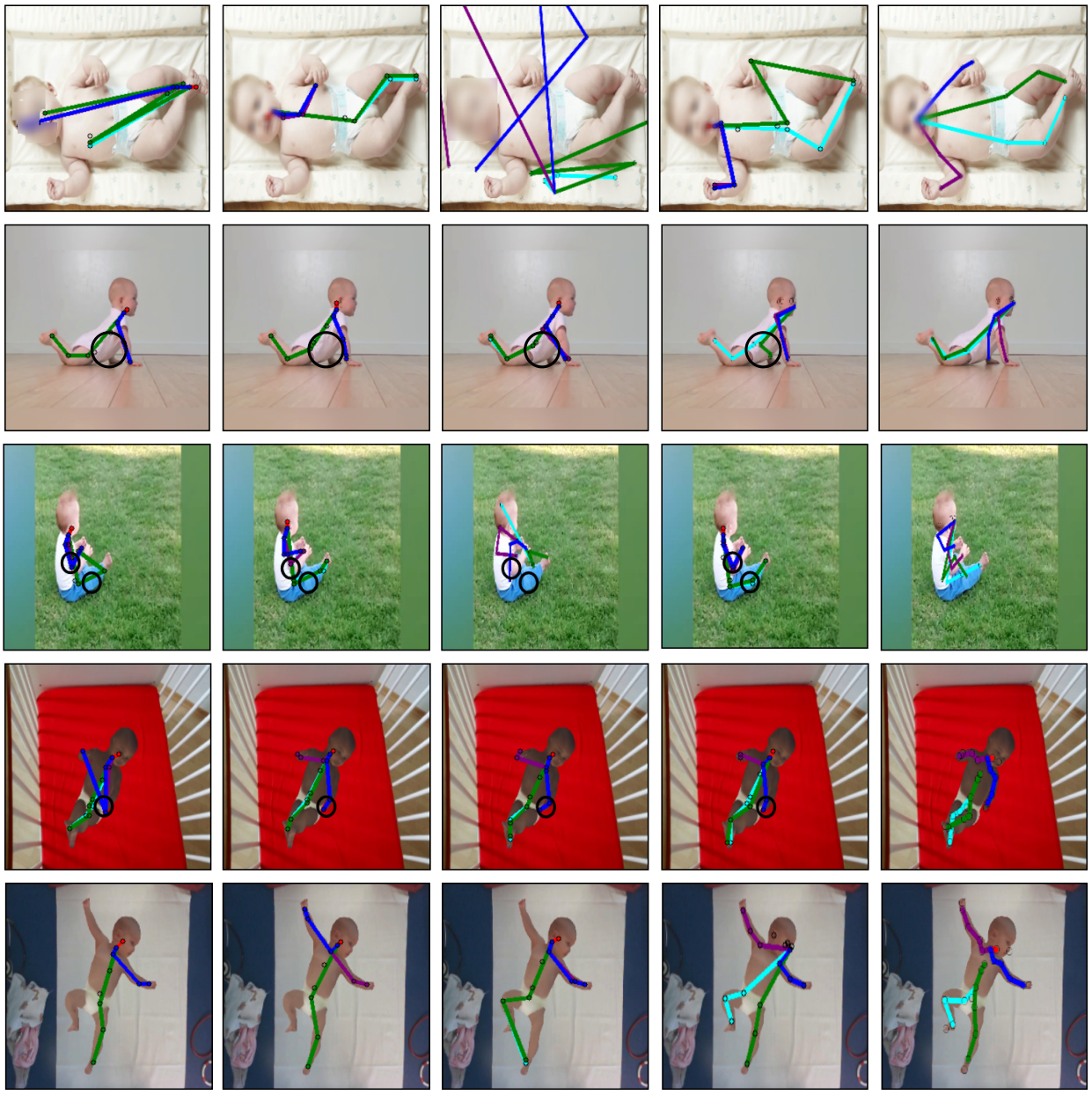

Figure 3: Qualitative results on SURREAL → SyRIP (top rows) and SURREAL → MINI-RGBD (bottom rows). The infant prior is essential for predicting plausible poses where baseline methods fail. ◯ denotes self-occluded regions correctly estimated by our method.

the Mini-RGBD infant dataset, the results of which are summarized in Table 3. Notably, despite the fully supervised manner by which FiDIP was trained on the SyRIP infant dataset, when evaluated on the MiniRGBD dataset it fails to resolve the domain gap. On the other hand, the annotation-free approach of `SHIFT` manages to successfully bridge the domain gap and outperform the FiDIP

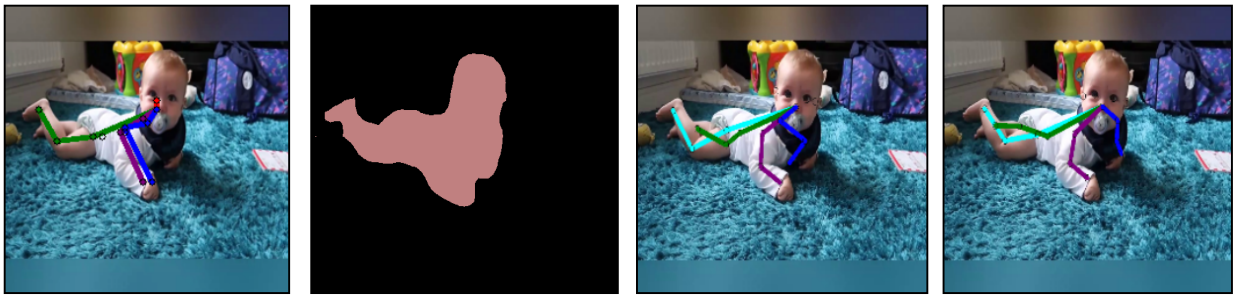

Figure 4: **Tackling Self-Occlusions: SURREAL → SyRIP.** UniFrame (left) fails to estimate the lower back and left hand, whereas SHIFT (middle) provides a reasonable estimate aligned with the segmentation mask and ground truth (right).

supervised infant pose estimation model.

4.3 Qualitative Evaluation of Self-Occlusion

We further highlight the self-occlusion mitigation of SHIFT in Figure 4. The UniFrame method fails to accurately evaluaten a sample exhibiting severe self occlusion. However, the SHIFT framework leverages the pose prior module and keypoint segmentation mapping module to effectively reason the anatomically plausible location of occluded joints. Specifically, it reasonably estimates the lower back and left hand of the infant subject, both of which are under heavy self-occlusion by the other body parts.

5 Conclusion

In all of these cases, the SHIFT framework clearly demonstrates robust performance in unsupervised domain adaptation from adult humans to infant humans without the use of annotations in the infant human dataset. By leveraging the mean teacher framework [23], SHIFT effectively performs self-supervised domain adaptation to bridge the anatomical domain gaps of adult and infant humans [13]. SHIFT accomplishes this by supplementing the baseline mean teacher framework with an anatomical pose prior module [24, 18] and a segmentation module [2, 3] to restrict and refine its predictions to an anatomically plausible distribution. Through this architecture, SHIFT successfully overcomes the severe self-occlusions which frequently occur in the infant pose distribution [8]. This accomplishment marks a significant milestone in the development of pose estimation

methodologies for contexts where the target domain has limited or restricted coverage, may not contain sufficient annotations for supervised learning methods, and presents additional anatomical challenges to the standard adult human pose distribution [10, 15].

Further implementation details supplementary to this work are detailed below.

6 Supplementary Material

Table 4: We analyse the effects of each loss term and module in this table for **SURREAL** [26] → **MINI-RGBD** [8].

Module	**Loss Terms**				**PCK@0.05**
	$\mathscr{L}_{\text{sup}}$	$\mathscr{L}_{\text{cons}}$	$\mathscr{L}_{p}$	$\mathscr{L}_{\text{ctx}}$	
Pre-Training	✓	✗	✗	✗	47.40
UniFrame [13]	✓	✓	✗	✗	51.50
UniFrame + Pose Prior	✓	✓	✓	✗	53.60
SHIFT	✓	✓	✓	✓	**56.40**

Table 5: We analyze the effects of different τ_c values on performance (**PCK@0.05**). **SURREAL** [26] is the source dataset.

τ_c	0.1	0.3	0.5	0.7	0.9
SyRIP [9]	35.00	39.00	**39.80**	37.50	35.10
Mini-RGBD [8]	53.00	53.70	**56.40**	54.10	53.50

Table 6: **Quantitative Results (PCK@0.05)** for the **Hybrid MiniRGBD-SyRIP** target domain, after completing 70 epochs of training. Evaluation uses an anatomical prior pre-trained on Human3.6M [11] and synthetic SyRIP samples.

Algorithm	SURREAL → Hybrid Target							
	Head	Sld.	Elb.	Wrist	Hip	Knee	Ankle	Avg.
Source only	85.80	84.40	76.80	76.20	76.00	65.90	74.40	75.60
Oracle	96.30	95.10	90.30	82.00	88.70	78.20	79.60	85.70
UniFrame [13]	85.50	74.20	88.20	**71.10**	31.10	**80.00**	77.10	70.30
SHIFT	**89.00**	**88.20**	**88.90**	66.50	**82.60**	75.60	76.00	**79.60**

6.1 Ablation Study: Module Impact

To evaluate the influence of each of the modules and their respective loss terms, we perform an ablation analysis when domain adapting from the SURREAL adult dataset [26] to the MiniRGBD

infant dataset [8]. Summarized in Table 4, the results demonstrate a step-wise improvement in performance with the cumulation of each module. Ultimately, the inclusion of both the anatomical pose prior module [24, 18] and the keypoint segmentation mapping module [3] significantly improve the overall performance of the SHIFT architecture.

6.2 Sensitivity Analysis: Pseudo-Label Threshold

We also evaluate the effect of the confidence threshold which we use for selecting the pseudo-labels from the teacher model to evaluate the student model. For this analysis, we perform domain adaptation from SURREAL to SyRIP [9] and SURREAL to MiniRGBD [8]. Summarized in Table 5, the results demonstrate that both high and low threshold values negatively impact the adaptation process, with a sweet spot of $\tau_c = 0.5$ resulting in the best performance. These results suggest that an overly low threshold degrades the quality of pseudo-labels while an overly high threshold filters out most pseudo-labels, thereby contaminating the evaluation step when training the student model [23, 13].

6.3 Hybrid MiniRGBD-SyRIP Dataset Curation

When comparing SHIFT to the FiDIP architecture [9], we noticed that domain adaptation from MiniRGBD to SyRIP produces unsatisfactory results for both methods. The MiniRGBD dataset contains a limited diversity of movements in each sample, due to the supine nature of the sampling setup for this dataset [8]. On the other hand, despite the smaller size of the SyRIP dataset, the diversity in infant pose distribution allows for more effective training, which is reflected in the resulting evaluation, previously discussed and summarized in Table 3 [9].

To effectively leverage the strengths of both of the infant human datasets, we curated a hybrid MiniRGBD-SyRIP dataset which leverages the larger size of the MiniRGBD dataset and the greater diversity of poses in the SyRIP dataset. The training split consists of ten thousand samples from the MiniRGBD dataset and 200 real infant images from the SyRIP dataset. We reserve the remaining two thousand MiniRGBD samples and the other three hundred real infant SyRIP samples for testing. We utilize the remaining one thousand synthetic infant samples from the SyRIP dataset in combination with an additional auxiliary adult human dataset, the Human3.6M dataset [11] to

train a new anatomical pose prior which effectively learns the general anatomy of human subjects and finetunes on the patterns of self-occlusion frequently observed in the infant pose distribution. Note that we split the SyRIP infant dataset into the synthetic and the real samples in order to avoid data leakage between the anatomical pose prior module training and the mean teacher adaptation training protocols [9, 24].

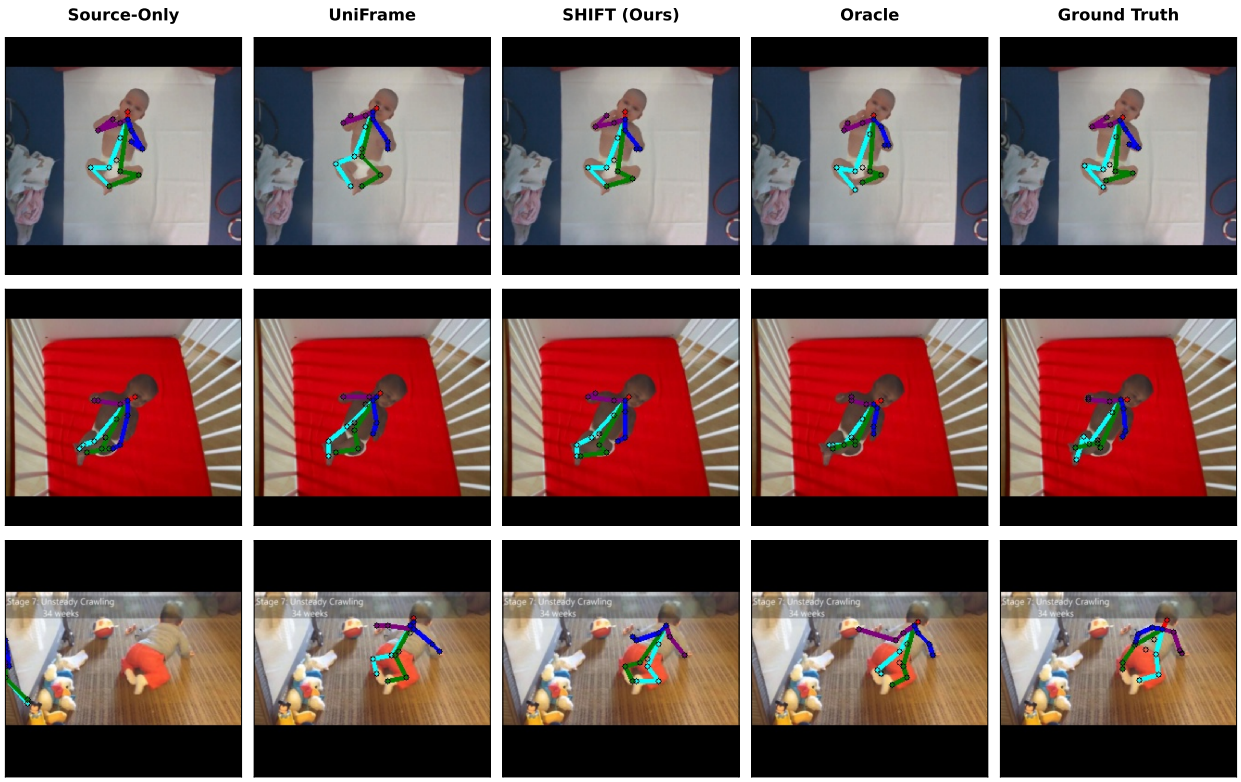

Figure 5: **Qualitative Comparison of Infant Pose Estimation across different domain adaptation strategies.** We visualize keypoint predictions on samples from the SyRIP [9] and MINI-RGBD [8] target domains. (Left to Right): The *Source-Only* model trained on the adult dataset (SURREAL [26]) fail to generalize to infant proportions. The *UniFrame* [13] baseline exhibits significant anatomical errors and joint collapse due to a lack of structural constraints. Our proposed method (`SHIFT`) consistently produces anatomically plausible poses which consistently closely align with the *Ground Truth*, and in some cases, perform better than the *Oracle* model, specifically in challenging regions such as the hips and self-occluded limbs.

6.4 Results on Hybrid Target Domains

Using this new hybrid infant dataset and hybrid adult-infant anatomical pose prior with the existing keypoint segmentation mapping module, we train `SHIFT` to adapt from the SURREAL synthetic adult human dataset [26] to the hybrid MiniRGBD-SyRIP real infant human dataset. Compared to

SHIFT adaptation to the individual infant datasets, this hybrid approach demonstrates a significant increase in performance and accuracy, summarized in Table 6 and visualized in Figure 5. In this scenario, SHIFT achieves a 9% increase over the baseline UniFrame [13] standard, despite accuracy drops in the wrist and knee joints. Qualitatively, in Figure 5 we observe that in samples of high self-occlusion and inter-limb interaction, the Source-Only [7] and the UniFrame [13] methods do not consistently output anatomically plausible poses. The Oracle model, despite extensive supervised training on the infant human dataset, also fails to accurately learn the infant pose distribution. In contrast, SHIFT demonstrates a robust domain adaptation, and successfully captures the anatomical distinctions of the infant pose distribution. These results suggest that a higher diversity and quantity of infant samples may benefit training sessions. Furthermore, the use of a hybrid synthetic adult and infant dataset to train the pose prior serves to reinforce general human anatomy while finetuning to the distinctly infant pose characteristics of severe self-occlusion.

References

[1] Sarosij Bose, Hannah Dela Cruz, Arindam Dutta, Elena Kokkoni, Konstantinos Karydis, and Amit K. Roy-Chowdhury. Leveraging synthetic adult datasets for unsupervised infant pose estimation, 2025. 3

[2] Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. Rethinking atrous convolution for semantic image segmentation, 2017. 13, 19

[3] Arindam Dutta, Rohit Lal, Dripta S Raychaudhuri, Calvin-Khang Ta, and Amit K Roy-Chowdhury. Poise: Pose guided human silhouette extraction under occlusions. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 6153–6163, 2024. 10, 13, 19, 21

[4] Riccardo Gatto and Sreenivasa Rao Jammalamadaka. The generalized von mises distribution. *Statistical Methodology*, 4(3):341–353, 2007. 12

[5] 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. 8

[6] Mijna Hadders-Algra. General movements: a window for early identification of children at high risk for developmental disorders. *The Journal of pediatrics*, 145(2):S12–S18, 2004. 8

[7] 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. 12, 16, 23

[8] 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. 8, 15, 16, 17, 19, 20, 21, 22

[9] Xiaofei Huang, Nihang Fu, Shuangjun Liu, and Sarah Ostadabbas. 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. 8, 15,

16, 17, 20, 21, 22

[10] 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. 7, 20

[11] 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. 5, 8, 20, 21

[12] 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. 5, 7, 9, 11, 12, 15, 16, 17

[13] 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. 5, 6, 7, 9, 11, 15, 16, 17, 19, 20, 21, 22, 23

[14] Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, 2009. 5

[15] Julienne 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 Humancentric AI (R2HCAI), Washington, DC*, 2023. 7, 8, 20

[16] Alejandro Newell, Kaiyu Yang, and Jia Deng. Stacked hourglass networks for human pose estimation, 2016. 6

[17] Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks, 2016. 14

[18] Dripta S Raychaudhuri, Calvin-Khang Ta, Arindam Dutta, Rohit Lal, and Amit K Roy-

Chowdhury. Prior-guided source-free domain adaptation for human pose estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14996–15006, 2023. 7, 8, 9, 12, 13, 17, 19, 21

[19] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115(3):211–252, 2015. 6

[20] 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. 8

[21] 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. 7

[22] 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. 6

[23] 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. 5, 6, 8, 9, 10, 12, 19, 21

[24] 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. 7, 9, 12, 13, 19, 21, 22

[25] 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. 10

[26] 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. 5, 8, 10, 14, 15, 17, 20, 22

[27] 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. 16