# Efficient Domain Adaptation via Generative Prior for 3D Infant Pose Estimation

Zhuoran Zhou[1] Zhongyu Jiang[1] Wenhao Chai[1]
Cheng-Yen Yang[1] Lei Li[2] Jenq-Neng Hwang[1]

[1] University of Washington [2] University of Copenhagen

{zhouz47, zyjiang, wchai, cycyang, hwang}@uw.edu, lilei@di.ku.dk

## Abstract

*Although 3D human pose estimation has gained impressive development in recent years, only a few works focus on infants, that have different bone lengths and also have limited data. Directly applying adult pose estimation models typically achieves low performance in the infant domain and suffers from out-of-distribution issues. Moreover, the limitation of infant pose data collection also heavily constrains the efficiency of learning-based models to lift 2D poses to 3D. To deal with the issues of small datasets, domain adaptation and data augmentation are commonly used techniques. Following this paradigm, we take advantage of an optimization-based method that utilizes generative priors to predict 3D infant keypoints from 2D keypoints without the need of large training data. We further apply a guided diffusion model to domain adapt 3D adult pose to infant pose to supplement small datasets. Besides, we also prove that our method, ZeDO-i, could attain efficient domain adaptation, even if only a small number of data is given. Quantitatively, we claim that our model attains state-of-the-art MPJPE performance of **43.6** mm on the SyRIP dataset and **21.2** mm on the MINI-RGBD dataset.*

## 1. Introduction

3D human pose estimation has been a popular research area these days. Similarly, pose estimation for infants plays an important role in risk assessment and healthcare monitoring [28]. However, due to privacy and the difficulty of data collection, public infant pose datasets are rare and limited, and manual labeling is unreliable and expensive. Therefore, it is challenging to train an efficient deep-learning model for infant pose estimation from scratch without sufficient data. To address this limitation, it is natural to think about transferring or tuning an existing adult-based pose estimation model on infant datasets to fully take advantage of similar kinetics of human body pose. Previous work like [11] tried to adapt a 2D adult pose detector to the infant domain,

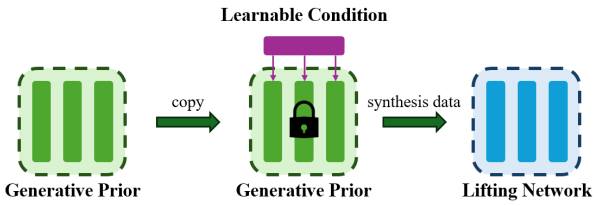


Figure 1. The overall flowchart of our method. Our model aims to adapt a generative prior pre-trained model based on large adult pose data to the infant domain via a controllable branch or fine-tuning. Then, we utilize generative prior in the infant domain to perform optimization work that predicts 3D infant pose from 2D.

but little work has been discussed for 3D infant pose estimation. Therefore, in this paper, we would like to pursue a method that can efficiently predict 3D infant poses even with small infant datasets, by taking advantage of the general kinetic knowledge transferred from an adult adult-based pre-trained model, as the flowchat shown in Figure 1.

Though learning-based 3D pose estimation models typically learn better features and obtain better performance compared to optimization algorithms, they inevitably require much more data in training for sufficient feature learning. Besides, the majority of human pose estimation learning models suffer from out-of-distribution issues, which make it hard for them to apply in practical scenarios or test data whose characteristics are far from the training data. However, this can easily happen for infant pose as few public datasets are available to support a more general 3D model's training, and cameras in hospitals or healthcare institutions may have different camera settings, leading to an unpredictable domain gap. Fortunately, previous optimization works [2, 32] are proved to be more insusceptible to distribution bias and robust in cross-domain tasks. Moreover, thanks to sophisticated 2D keypoint detectors, two-stage lifting networks are generally of higher accuracy than

one-stage networks, which directly predict 3D pose from raw images. In addition, we believe that generation models can be easily trained with few data and be domain-adapted efficiently compared to classic deep-learning models [26]. Therefore, inspired by ZeDO [16], we choose to apply a two-stage optimization-based method, named ZeDO-i, to address the lack of data and out-of-distribution issues under the assistance of generative priors. Given 2D keypoints, our model can iteratively adjust noisy 3D prediction under the constraint of 2D-3D projection and prior distribution learned. As we expected, generative priors learned in the adult domain could be effectively transferred to the infant domain without requiring a lot of data, and the optimization process can cope with challenging test data in reality. Moreover, to simulate the extreme condition of lack of data in small datasets, we also test our model with only 20 and 100 data during adaption and successfully validate our model's ability for efficient domain adaptation. Furthermore, we also introduce a guided diffusion model, which aims to supplement datasets by adapting adult pose to infant pose in order to address data limitation issues and reinforcement diversity. Finally, our method obtains SOTA performance in terms of MPJPE on two infant pose datasets. In this paper, we make the following contributions:

- We propose an optimization-based method using generative priors for 3D infant pose estimation. We attain SOTA performance on MINI-RGBD [7] and SyRIP [11]. We also claim that our model can achieve efficient domain adaptation even with a small number of data.

- We introduce a condition-guided diffusion model which can adapt adult human keypoints to similar infant keypoints for data augmentation purposes and further enhance performance.

## 2. Related Work

### 2.1. 3D Human Pose Estimation

3D Human Pose Estimation is one of the fundamental tasks in computer vision and is crucial to many downstream tasks, including Human Tracking [1, 34], Action Recognition [5, 29, 36, 41], Motion and Gait Analyses [10, 15, 40], and so on. There are three main approaches to realizing the 3D human pose estimation: optimization-based, 2D-3D lifting [20, 38], and image-based methods.

Optimization-based methods are not limited by any training dataset and are good at in-the-wild inference. However, the performance of previous optimization-based methods [2, 25, 33] is commonly worse than the performance of training-based networks. 2D-3D lifting methods follow a two-stage pipeline requiring a separate 2D human pose estimation model and a lifting network to map 2D human poses to 3D human poses in single frames or short sequences. Pavllo *et al*. [27] apply dilated temporal convolution to enhance 3D pose estimation for unlabeled videos in a semi-supervised method. Zhao *et al*. [39] design a novel graph convolution and take advantage of a graph convolution network (GCN) to learn inter-joint features and local and global relationships in a structured graph. On the other hand, image-based methods focus on directly regressing 3D human poses from RGB images. Kolotouros *et al*. [18] introduce SPIN (SMPL oPtimization IN the loop) by using a CNN to extract features from a cropped-out human image and regress the SMPL [22] parameters with the help of an optimization-based pose estimation pipeline to conduct semi-supervised learning. However, all the learning-based methods suffer from the use of small datasets in Infant Pose Estimation tasks. In this paper, we focus on how to conduct 3D infant pose estimation with limited data.

### 2.2. Infant Pose Estimation

Infant pose estimation, which aims at predicting 2D and 3D keypoints of infants in image and world coordinates, can lead to useful downstream tasks such as infant action recognition [12,36] or motion analysis [4,13]. Hesse *et al*. [7] are the first to present the MINI-RGBD dataset, which enables the experiment on 2D infant pose estimation. Subsequently, Huang *et al*. [11] propose a hybrid synthetic and real infant pose (SyRIP) dataset based on SMIL [24] with annotated 2D keypoints. Following the 2D infant pose estimation, the mainstream of 3D infant pose estimation works on RGB-D data. Wu *et al*. [35] measure infant movements by combining 2D keypoints and matching depth images collected by Kinect. Li [19] continues using the same pipeline but correcting depth information for a better matching between image and depth. However, Kinect may cause depth ambiguity if joints are occluded, and depth images are not always available in the infant monitor system. In [6], the author uses a 2D pose estimation model and a 3D lifting network pretrained on the adult dataset and fine-tuned on the infant dataset. Though this model achieves rather good performance on the MINI-RGBD dataset, it is basically learning-based and hard to adapt to more realistic data due to the domain gap. From our experience in human pose estimation, predicting 3D keypoints from 2D keypoint detection is easier than the one-stage method predicting 3D joints directly from raw images.

## 3. Methods

Our model primarily consists of a diffusion model to learn the prior, and an optimization algorithm to iteratively adjust 3D pose prediction. Additionally, we apply a condition-guided diffusion model for pose data augmentation. We demonstrate the method as followings: back-

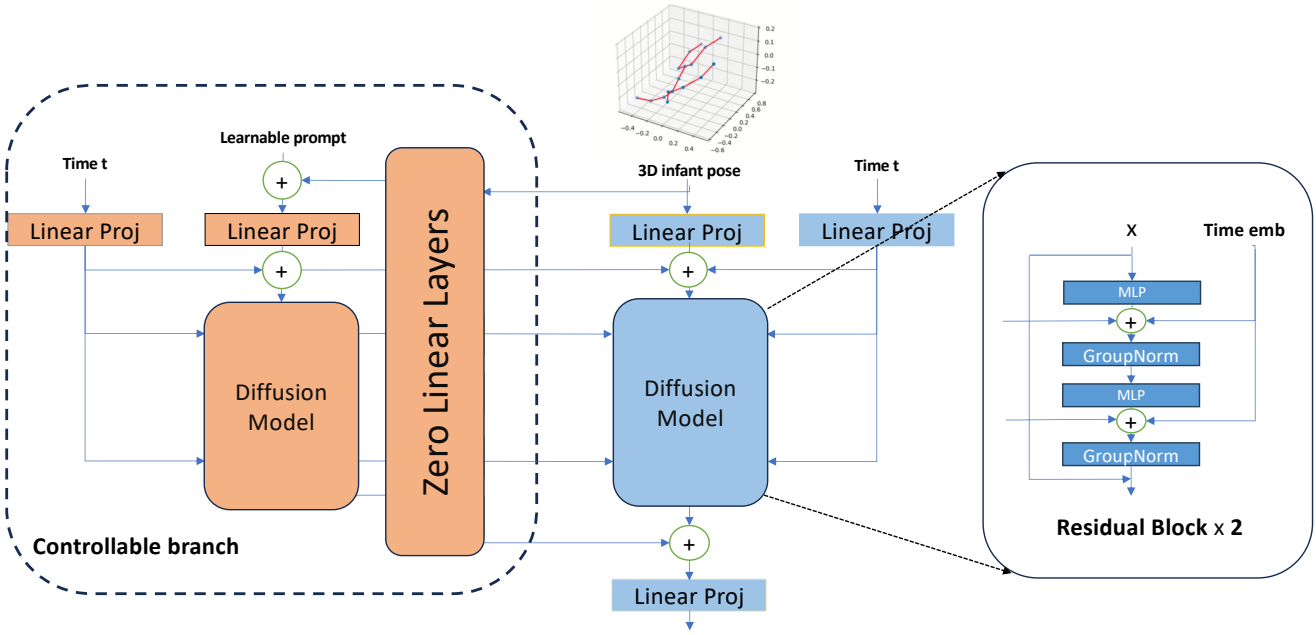


Figure 2. **Model architecture**. The modules excluded by the dotted box comprise our proposed score-matching-based prior learning model, and the modules inside the dotted box are the controllable branch used in one of our adaptation strategies. The prior learning model consists of MLP layers and two residual blocks displayed on the right side. The controllable branch takes a learning prompt as inputs, copies the weights from the prior model and only updates the copied weights during adaptation, while the original prior learning model is kept frozen. For convenience, we paint all frozen layers in controllable adaptation as blue and all updated layers as brown.

ground of diffusion model in section 3.1, generative prior model in section 3.2, optimization algorithm in 3.3, its controllable adaptation variant in 3.4 and condition-guided data augmentation in 3.5.

### 3.1. Preliminaries of Diffusion Model

Before introducing our diffusion model as the generative prior learning model, we briefly discuss its background for clarity. Diffusion models [9, 31] are popular generation models used in tasks like image generation [30], image inpainting [23], editing [3] and so on. During training, the diffusion model iteratively adds Gaussian noise to an image relative to a timestamp $t$ to the inputs and tries to recover inputs from a noise image in the reverse process. In this paper, we use Score-Matching-Network(SMN) [31] as our prior learner. SMN aims to train a score network $s_\theta(x)$ to approximate gradients of log probability of a score function $p_\theta(x)$, expressed as $s_\theta(x) \approx \nabla_x \log p_\theta(x)$, so the loss is generally represented as

$$E_{p(x)} \| \nabla_x \log p_\theta(x) - s_\theta(x) \|_2^2. \tag{1}$$

### 3.2. Infant Pose Prior Model

For pose estimation tasks on small datasets, learning-based deep-learning models suffer from out-of-distribution issues and insufficient resources to extract reliable features. Built upon the work of ZeDO [16], we also propose to use an optimization-based method to predict 3D keypoints from 2D keypoints along with a score-matching network diffusion model(SMN) [31] as our prior learner. Our final architecture is illustrated in 2. The modules excluded by the dotted bounding box comprise our proposed prior learning model, which takes root-relative infant keypoints, sized $B \times J \times 3$, and randomly sampled noise timestamp $t$ as inputs, where $J$ is the number of joints. The embedding layers are simple linear projection layers, which lift input dimension to $B \times 1024$, and then sum them up. Further, the embedding goes through the Score-Matching-Network diffusion model consisting of two residual blocks as backbones. Each of the residual blocks contains two residual-connected MLPs. The last output projection layer projects the feature back to pose joints. With the generative priors, our method can denoise a noisy 3D pose in the optimization stage if it violates the kinematic rules of infant poses.

### 3.3. Optimization Algorithm

Given a 2D infant pose and the intrinsic parameter, ZeDO-i first tries to compute the ray vectors emitted by the camera and initializes the predicted 3D keypoints on the rays to minimize 3D-2D projection errors. Further it activates the generation model to adjust the noisy 3D pose prediction based on its prior knowledge. After each ad-

justment, 3D keypoints may be off the rays, and the model again moves them onto the rays in the shortest distance. Our method runs this iteration 1000 times to iteratively achieve a reasonable 3D pose under a 3D-2D projection constraint. In experiments, we find that a pseudo intrinsic parameter which has a focal length of 2000 and a camera center equal to the image center also functions so one could apply it in practical cases.

In details, we first define ray vectors $\hat{V_{ray}}$ emitted from the camera using 2D keypoints $X_{2D}$ and real or pseudo intrinsic parameter $K$, in which focal length is always 2000 and the principal point is the image center point. Then we randomly choose a training 3D pose $X_{3D_init}$ and use an Adam optimizer [17], which helps us find an appropriate rotation $R_o$ and translation $T_0$ such that $\|K(R_0 X_{3D_i nit} + T_0) - X_{2d}\|_2$ is minimized. With $T_0$ known, we set all initial 3D keypoints on the rays with depth equal to $T_0$, and supposedly this 3D pose has zero projection error with the 2D ground truth. Next, we start $T = 1000$ times of optimization steps in which we first move 3D kyepoints to corresponding rays in the shortest distance if they are off the rays and then the prior model is used to adjust the noisy pose based on the prior distribution it learns.

In evaluation, we find that a noise level $t = (0, 0.1]$ works the best, and we also observe that performance is heavily dependent on the initial depth distance assigned. As the training data are all root-relative, the prior model may cause depth ambiguity if we don't limit depth in the first few steps. In practice, we get the lowest error when forcing the depth $T$ unchanged in the first 950 iterations and opening the constraint in the remaining 50.

### 3.4. Controlling Branch for Domain Adaptation

As the kinematics of infants and adults are similar, transferring a pre-trained adult pose model to the infant domain would intuitively boost the performance. Considering that directly fine-tuning a model trained on a huge amount of data to a small dataset may lead to overfitting, we propose a method inspired by Control-Net [37] to manipulate the adaption process of the generative priors. As shown in the left-side dotted box in Figure 2, we duplicate the weights of the prior model to the controllable branch. Like how Control-Net sets the condition, We set a learnable prompt with the same size as 3D pose as the controlling inputs and connect the prior model and controllable branch with a few zero linear layers, which are fully initialized as zero weights. Then internal embeddings are added back to the prior model before and after every residual block. During adaptation, all layers of the prior model have to be kept frozen, and only the controllable branch is open to weight update. In the experiment section, we will compare its performance with two other adaption strategies: fine-tuning a pre-trained adult prior model and training a prior model in the infant domain only.

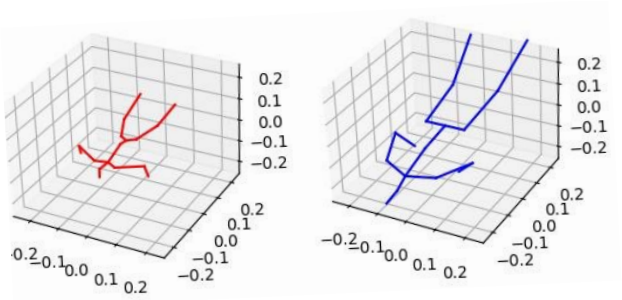

Figure 3. Left: augmented infant pose. Right: h36m adult pose. Our augmentation model converts the adult pose to a similar infant pose by adjusting scales and kinematic features like bone length without altering actions much.

### 3.5. Condition Guided Diffusion Model

If the only available data are too few to be used, one could convert the resourceful adult poses to an infant-like pose as data augmentation. To fit the augmented data to the kinematics of infants, we trained a score matching network diffusion model which takes in both adult and infant 3D poses along with two $1 \times 1000$ sized learnable condition tokens to represent whether the pose belongs to adults or infants. The architecture is similar to the prior model we used. We hope that the diffusion model would implicitly learn features like bone lengths and bone angles for two different domains and know how to distinguish their distribution. During inference, we ask the diffusion model to generate corresponding infant poses based on the given input adult poses, so the model adjusts the scale and angle according to the implicit knowledge of the pose prior, yet still keeps the pose semantic meanings, such as actions, as shown in Figure 3. We prove that adding these challenging poses enhances diversity in the ablation study.

## 4. Experiments

### 4.1. Datasets

We conduct our experiments on MINI-RGBD [7] and SyRIP [11], two public infant datasets with 2D-3D pose pairs. For pre-trained adult prior model and condition-guided diffusion model, we take advantage of Human3.6m [14].

**MINI-RGBD** includes 12 sequences of data, in total 12000 synthetic infant images, and also provides their 25 joint 2D and 3D keypoint pairs. We train on the first 9 sequences and test on the rest 3, following the 17-keypoint definition of Human3.6m.

**SyRIP** includes a diverse set of 700 real and 1000 synthetic infant images, generated by fitting SMIL [24] models to real images, supplemented with additional variants to the

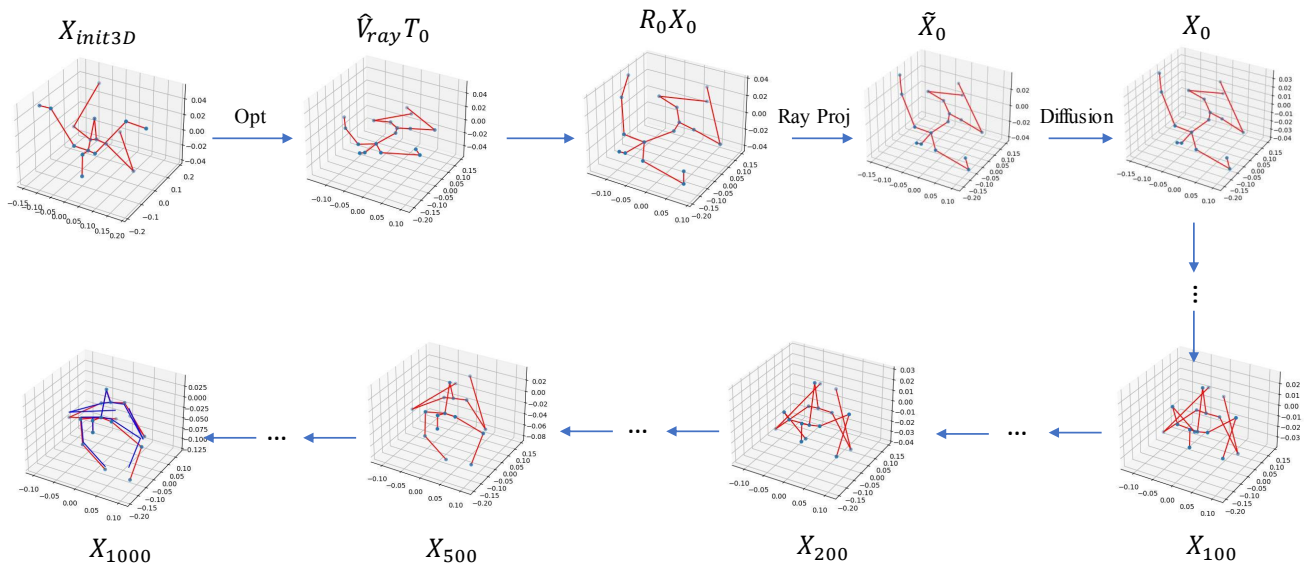


Figure 4. Visualization of the optimization algorithm. $\hat{V_{ray}}$ ray vectors are first calculated. $T_0$ and $R_0$ are then found via an optimizer, so the initial pose is $T_0$ in depth on the rays. We then run the optimization algorithm 1000 times. In each step, keypoints are moved toward the rays, represented by $\tilde{X_0}$ in the figure, and are also sent to the diffusion model to adjust the pose to get $X_0$. In the last step $X_{1000}$ the ground truth in blue is quite close to our prediction in red.

SMIL shape and pose parameters. Later, in total of 700 weak ground truth of 3D keypoints in [13] were manually corrected and made available. In this paper, we also train on their 600 weak ground truth 3D labels and test on the 100 real images.

**Human3.6m (H36m)** [14] is a single-frame 3D human pose benchmark, containing about 3.6 million 2D-3D human pose pairs. This dataset was collected in an indoor setting, consisting of 17 various actions. As this adult 3D pose dataset includes more actions and diversity than infant 3D poses, we intend to transfer its 3D poses to the infant domain for data augmentation.

### 4.2. Implementation Details

We pre-train our adult prior model on Human3.6m [14], which includes millions of adult 3D pose data, and further train all three adaptation strategies for 5000 epochs with a learning rate of $2 \times 10^{-4}$. During training, we set the total diffusion step as 1000 with a uniform noise level of $[0, 1.0]$. We use the Adam Optimizer with a batch size of 5000.

In inference, we choose a noise level $t$ in $(0, 0.1]$ and run the optimization for 1000 iterations. We keep the depth distance unchanged in the first 950 iterations.

The guided diffusion data augmentation diffusion model shares the same training configuration as the prior model. We choose a noise level in the range $(0, 1.0]$ and only run 100 iterations for the diversity of augmented data. In experiments, we add 600 augmented data to SyRIP and 4000 to MINI-RGBD.

| Methods | MPJPE (↓) |
|---|---|
| Kolotouros *et al*. [18] | 105.8 |
| Liu *et al*. [21] | 97.2 |
| Liu *et al*. (Finetuned) | 78.3 |
| ZeDO-i (GT) | **43.6** |
| ZeDO-i (DT) | 47.7 |

Table 1. 3D infant pose estimation results on SyRIP dataset under 12 joints setting. For a fair comparison, we list the performance of ZeDO-i with both estimated 2D keypoints and ground truth 2D keypoints.Estimated 2D keypoints and other method performance are provided in [13].

### 4.3. Experiment Results

In this section, we first compare our method's results to the previous SOTA in terms of MPJPE. In addition, we also test if our method can be efficiently adapted to the infant domain with 20 and 100 data only in order to simulate extreme situations. Further, we evaluate all domain adaptation strategies of our model including the controllable adaptation method (CA), fine-tuning from the adult-based diffusion model (FT), and training from scratch on infant data to seek the best adaptation approach.

### 4.4. Results on SyRIP

Similar to previous works, we have the same training and testing sets as [13] with only 12 keypoints of limbs for fair

| Methods | MPJPE (↓) |
|---|---|
| Hesse *et al.* [8] | 44.9 |
| Ellershaw* *et al.* [6] | 34.2 |
| Ellershaw *et al.* | 28.5 |
| ZeDO-i | **21.2** |

Table 2. 3D infant pose estimation results on MINI-RGBD under 16 joints setting. We list the best performance among the three strategies. * denotes w/o adult pre-training. We evaluate 16 keypoints to keep aligned with the setting of the previous SOTA.

| Datasets | CA | FT | From Scratch |
|---|---|---|---|
| SyRIP ($S$=20) | **67.8** | 69.4 | 72.3 |
| SyRIP ($S$=100) | **56.4** | 60.8 | 60.6 |
| SyRIP(GT) | 49.4 | **47.7** | 54.0 |
| SyRIP (augmented) | 45.5 | **43.6** | 48.9 |

Table 3. MPJPE performance of different strategies on SyRIP. The controllable adaptation approach achieves better performance than the other two approaches when the data number is small.

comparison. Observed from Table 1, our method clearly achieves the SOTA performance even with only 20 training data. Moreover, as shown in Table 3, we observe that the controllable adaptation approach achieves better results than fine-tuning when the data number is small, therefore controllable adaptation is more suitable for limited data in such more practical and diverse scenarios. Both adaptation and fine-tuning from the adult domain are better than training from scratch, indicating that knowledge from the adult domain is necessary.

### 4.5. Results on MINI-RGBD

For a fair comparison with previous works [6], we follow their keypoint definitions and show the results in Table 2. Our method beats all previous SOTA to a great extent.
Besides, as shown in Table 4, direct fine-tuning adult pre-trained model on MINI-RGBD attains lower error, which is different from SyRIP. We suspect that the discrepancy between training and testing sets leads to this observation, as MINI-RGBD is full of synthetic images with rather less discrepancy compared to the SyRIP dataset. Moreover, we also include the results of 16 keypoints like the previous SOTA, showing that our model is already comparable to the previous SOTA with only 100 training data.

## 5. Ablation Studies

### 5.1. Data Augmentation Diversity

In this section, we demonstrate how our condition-guided data augmentation method enhances the diversity in the MINI-RGBD dataset as its data are all synthesized and narrow-distributed. We analyze bone lengths and bone angles of the original dataset and our augmented data. As shown in Figure 5, we randomly choose one bone and compare their lengths. Our augmented bone length spans over a wider range of scales. Similarly, we show comparisons of bone angles in Figure 6 and get the same conclusion. Tables 4 and 3 also justify this conclusion quantitatively.

| Datasets | CA | FT | From Scratch | Best($J$ = 16) |
|---|---|---|---|---|
| MINI-RGBD ($S$=20) | 38.7 | 36.8 | **36.4** | 34.6 |
| MINI-RGBD ($S$=100) | 34.8 | **31.9** | 33.7 | 29.4 |
| MINI-RGBD Full | 25.5 | **24.1** | 27.4 | 22.8 |
| MINI-RGBD (Augmented) | 20.7 | **19.9** | 21.3 | 21.2 |

Table 4. MPJPE performance of different strategies on MINI-RGBD. Here, we not only evaluate all 17 keypoints according to H36M's keypoint definition but also list their performance in 16 keypoints for the convenience of fair comparison with previous works.

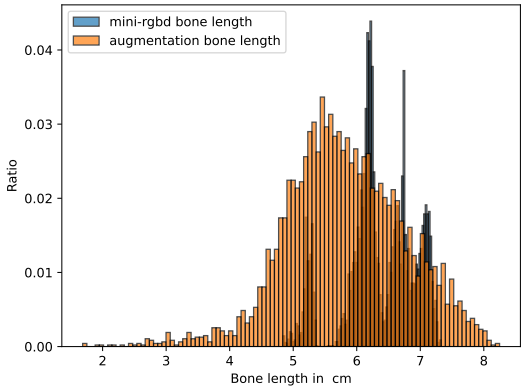


Figure 5. The data distribution of bone length in the augmented dataset is better than the original MINI-RGBD. Our augmented dataset spans over a wider range of bone lengths.

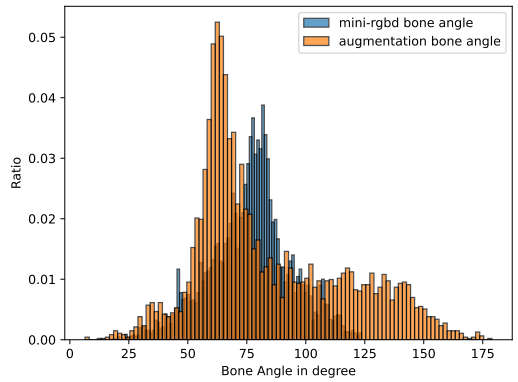


Figure 6. The data distribution of bone angle in the augmented dataset is better than the original MINI-RGBD. Our augmented data have a wider range of bone angles.

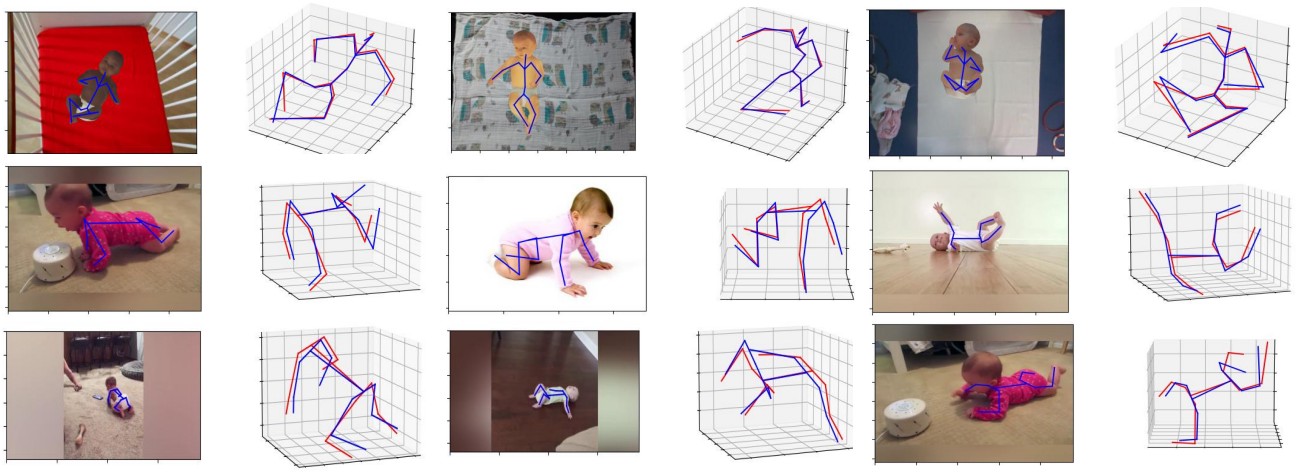

Figure 7. Visualizations of 2D ground truths and our 3D predictions on MINI-RGBD(Top Line) and SyRIP(Middle and Bottom Line). Our 3D predictions are colored in red, and the ground truth are in blue. The shapes and poses in general are well aligned.

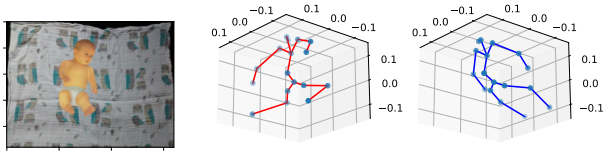

Figure 8. One of the failure examples achieves the highest MPJPE of 160mm in MINI-RGBD. Left side: our prediction. Right side: ground truth.

| Datasets | ZeDO-i | VideoPose3D [27] |
|---|---|---|
| SyRIP | **49.6** | 145.7 |
| MINI-RGBD | **27.4** | 106.7 |

Table 5. MPJPE performance of different pose estimation models trained from scratch without augmentation. Clearly, ZeDO-i is more suitable for infant small dataset than learning-based models.

### 5.2. Comparison with 3D pose estimation model

To prove the efficiency of our two-stage optimization method, we compare its performance with other classic learning-based 3D pose estimation models widely applied in human pose estimation tasks. We train all the models from scratch on infant datasets without data augmentation.

As shown in Table 5, our method outperforms the classic 2D-3D lifting human pose estimation model, VideoPose3D [27], which further proves our claim that the proposed optimization method can better fit the task of small-dataset domain adaption in 3D pose estimation than other learning-based models.

## 6. Limitation

Though our method achieves impressive performance in small datasets like infant 3D pose, it still needs accurate 2D keypoints. Additionally, the prediction results of our method also depend on the depth distance $T_0$ defined in initialization since we find that the generation model only learns root-relative priors with little knowledge of spatial depth. Besides, like all optimization works aiming to minimize 2D-3D projection error, our method may also suffer from one-to-many mappings. For example, we show one failure example in MINI-RGBD in Figure 8. Here our model fails to predict the correct $T$ of 3D keypoints in spite of the matched 2D projections. We calculated the median MPJPE error which is 4mm lower than the mean, which implies that these extreme outlines are very rare.

## 7. Conclusion

We propose an optimization method which applies generative priors of the infant pose to predict 3D infant keypoints. We show that our method achieves SOTA on MINI-RGBD and SyRIP and attains efficient domain adaptation using a small amount of data. Besides, we compare three training strategies for our model, in which fine-tuning an adult pre-trained generative model seems more efficient for MINI-RGBD and the whole SyRIP dataset, but the controllable adaptation version performs better in SyRIP when only 20 and 100 data are available. We also introduce a condition-guided diffusion model which enhances the kinematic diversity and boosts overall results. In general, we state that our method fits the small-dataset 3D infant pose estimation very well and attains outstanding performance.

## References


[1] Mykhaylo Andriluka, Umar Iqbal, Eldar Insafutdinov, Leonid Pishchulin, Anton Milan, Juergen Gall, and Bernt Schiele. Posetrack: A benchmark for human pose estimation and tracking. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5167–5176, 2018. 2

[2] Federica Bogo, Angjoo Kanazawa, Christoph Lassner, Peter Gehler, Javier Romero, and Michael J Black. Keep it smpl: Automatic estimation of 3d human pose and shape from a single image. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part V 14*, pages 561–578. Springer, 2016. 1, 2

[3] Shidong Cao, Wenhao Chai, Shengyu Hao, Yanting Zhang, Hangyue Chen, and Gaoang Wang. Difffashion: Reference-based fashion design with structure-aware transfer by diffusion models. *arXiv preprint arXiv:2302.06826*, 2023. 3

[4] Claire Chambers, Nidhi Seethapathi, Rachit Saluja, Helen Loeb, Samuel R Pierce, Daniel K Bogen, Laura Prosser, Michelle J Johnson, and Konrad P Kording. Computer vision to automatically assess infant neuromotor risk. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 28(11):2431–2442, 2020. 2

[5] Haodong Duan, Yue Zhao, Kai Chen, Dahua Lin, and Bo Dai. Revisiting skeleton-based action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2969–2978, 2022. 2

[6] Simon Ellershaw, Luca Schmidtke, Nidal Khatib, Jonathan Eden, Sofia Dall'Orso, Silvia Muceli, Etienne Burdet, Niamh Nowlan, Tomoki Arichi, and Bernhard Kainz. 3d infant pose estimation using transfer learning. 2, 6

[7] Nikolas Hesse, Christoph Bodensteiner, Michael Arens, Ulrich Hofmann, Raphael Weinberger, and Andreas Schroeder. *Computer Vision for Medical Infant Motion Analysis: State of the Art and RGB-D Data Set: Munich, Germany, September 8-14, 2018, Proceedings, Part VI*, pages 32–49. 01 2019. 2, 4

[8] Nikolas Hesse, Gregor Stachowiak, Timo Breuer, and Michael Arens. Estimating body pose of infants in depth images using random ferns. In *2015 IEEE International Conference on Computer Vision Workshop (ICCVW)*, pages 427–435, 2015. 6

[9] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 3

[10] Nicholas Howe, Michael Leventon, and William Freeman. Bayesian reconstruction of 3d human motion from single-camera video. *Advances in neural information processing systems*, 12, 1999. 2

[11] 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. 1, 2, 4

[12] Xiaofei Huang, Lingfei Luan, Elaheh Hatamimajoumerd, Michael Wan, Pooria Daneshvar Kakhaki, Rita Obeid, and Sarah Ostadabbas. Posture-based infant action recognition in the wild with very limited data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4911–4920, 2023. 2

[13] Xiaofei Huang, Michael Wan, Lingfei Luan, Bethany Tunik, and Sarah Ostadabbas. Computer vision to the rescue: Infant postural symmetry estimation from incongruent annotations. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1909–1917, 2023. 2, 5

[14] 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. 4, 5

[15] Zhongyu Jiang, Haorui Ji, Samuel Menaker, and Jenq-Neng Hwang. Golfpose: Golf swing analyses with a monocular camera based human pose estimation. In *2022 IEEE International Conference on Multimedia and Expo Workshops (ICMEW)*, pages 1–6. IEEE, 2022. 2

[16] Zhongyu Jiang, Zhuoran Zhou, Lei Li, Wenhao Chai, Cheng-Yen Yang, and Jenq-Neng Hwang. Back to optimization: Diffusion-based zero-shot 3d human pose estimation. *arXiv preprint arXiv:2307.03833*, 2023. 2, 3

[17] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 4

[18] Nikos Kolotouros, Georgios Pavlakos, Michael J Black, and Kostas Daniilidis. Learning to reconstruct 3d human pose and shape via model-fitting in the loop. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 2252–2261, 2019. 2, 5

[19] Min Li, Fan Wei, Yu Li, Sicong Zhang, and Guanghua Xu. Three-dimensional pose estimation of infants lying supine using data from a kinect sensor with low training cost. *IEEE Sensors Journal*, 21(5):6904–6913, 2020. 2

[20] Hanbing Liu, Jun-Yan He, Zhi-Qi Cheng, Wangmeng Xiang, Qize Yang, Wenhao Chai, Gaoang Wang, Xu Bao, Bin Luo, Yifeng Geng, et al. Posynda: Multi-hypothesis pose synthesis domain adaptation for robust 3d human pose estimation. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 5542–5551, 2023. 2

[21] Shuangjun Liu, Xiaofei Huang, Nihang Fu, and Sarah Ostadabbas. Heuristic weakly supervised 3d human pose estimation in novel contexts without any 3d pose ground truth. *arXiv preprint arXiv:2105.10996*, 2021. 5

[22] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J Black. Smpl: A skinned multi-person linear model. *ACM transactions on graphics (TOG)*, 34(6):1–16, 2015. 2

[23] Andreas Lugmayr, Martin Danelljan, Andres Romero, Fisher Yu, Radu Timofte, and Luc Van Gool. Repaint: Inpainting using denoising diffusion probabilistic models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11461–11471, 2022. 3

[24] Mengmeng Ma, Jian Ren, Long Zhao, Sergey Tulyakov, Cathy Wu, and Xi Peng. Smil: Multimodal learning with

severely missing modality. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 2302–2310, 2021. 2, 4

[25] Lea Muller, Ahmed AA Osman, Siyu Tang, Chun-Hao P Huang, and Michael J Black. On self-contact and human pose. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9990–9999, 2021. 2

[26] Atsuhiro Noguchi and Tatsuya Harada. Image generation from small datasets via batch statistics adaptation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2750–2758, 2019. 2

[27] Dario Pavllo, Christoph Feichtenhofer, David Grangier, and Michael Auli. 3d human pose estimation in video with temporal convolutions and semi-supervised training. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7753–7762, 2019. 2, 7

[28] Véronique Pierrat, Laetitia Marchand-Martin, Stéphane Marret, Catherine Arnaud, Valérie Benhammou, Gilles Cambonie, Thierry Debillon, Marie-Noëlle Dufourg, Catherine Gire, François Goffinet, Monique Kaminski, Alexandre Lapillonne, Andrei Scott Morgan, Jean-Christophe Rozé, Sabrina Twilhaar, Marie-Aline Charles, and Pierre-Yves Ancel. Neurodevelopmental outcomes at age 5 among children born preterm: Epipage-2 cohort study. *BMJ*, 373, 2021. 1

[29] Zhenting Qi, Ruike Zhu, Zheyu Fu, Wenhao Chai, and Volodymyr Kindratenko. Weakly supervised two-stage training scheme for deep video fight detection model. In *2022 IEEE 34th International Conference on Tools with Artificial Intelligence (ICTAI)*, pages 677–685. IEEE, 2022. 2

[30] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *International Conference on Machine Learning*, pages 8821–8831. PMLR, 2021. 3

[31] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020. 3

[32] Zheng Tang, Renshu Gu, and Jenq-Neng Hwang. Joint multi-view people tracking and pose estimation for 3d scene reconstruction. In *2018 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6. IEEE, 2018. 1

[33] Zheng Tang, Renshu Gu, and Jenq-Neng Hwang. Joint multi-view people tracking and pose estimation for 3d scene reconstruction. In *2018 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6. IEEE, 2018. 2

[34] Manchen Wang, Joseph Tighe, and Davide Modolo. Combining detection and tracking for human pose estimation in videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11088–11096, 2020. 2

[35] Qingqiang Wu, Guanghua Xu, Fan Wei, Longting Chen, and Sicong Zhang. Rgb-d videos-based early prediction of infant cerebral palsy via general movements complexity. *IEEE Access*, 9:42314–42324, 2021. 2

[36] Cheng-Yen Yang, Zhongyu Jiang, Shih-Yu Gu, Jenq-Neng Hwang, and Jang-Hee Yoo. Unsupervised domain adaptation learning for hierarchical infant pose recognition with synthetic data. In *2022 IEEE International Conference on Multimedia and Expo (ICME)*, pages 01–06. IEEE, 2022. 2

[37] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3836–3847, 2023. 4

[38] Zhenyu Zhang, Wenhao Chai, Zhongyu Jiang, Tian Ye, Mingli Song, Jenq-Neng Hwang, and Gaoang Wang. Mpm: A unified 2d-3d human pose representation via masked pose modeling. *arXiv preprint arXiv:2306.17201*, 2023. 2

[39] Long Zhao, Xi Peng, Yu Tian, Mubbasir Kapadia, and Dimitris N Metaxas. Semantic graph convolutional networks for 3d human pose regression. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3425–3435, 2019. 2

[40] Zhonghan Zhao, Wenhao Chai, Shengyu Hao, Wenhao Hu, Guanhong Wang, Shidong Cao, Mingli Song, Jenq-Neng Hwang, and Gaoang Wang. A survey of deep learning in sports applications: Perception, comprehension, and decision. *arXiv preprint arXiv:2307.03353*, 2023. 2

[41] Jianxiong Zhou, Zhongyu Jiang, Jang-Hee Yoo, and Jenq-Neng Hwang. Hierarchical pose classification for infant action analysis and mental development assessment. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1340–1344. IEEE, 2021. 2