

Invariant Representation Learning for Infant Pose Estimation with Small Data

Xiaofei Huang, Nihang Fu, Shuangjun Liu, Sarah Ostadabbas
 Augmented Cognition Lab, Electrical and Computer Engineering Department,
 Northeastern University, Boston, MA, USA

Abstract— Infant motion analysis is a topic with critical importance in early childhood development studies. However, while the applications of human pose estimation have become more and more broad, models trained on large-scale adult pose datasets are barely successful in estimating infant poses due to the significant differences in their body ratio and the versatility of their poses. Moreover, the privacy and security considerations hinder the availability of adequate infant pose data required for training of a robust model from scratch. To address this problem, this paper presents (1) building and publicly releasing a hybrid synthetic and real infant pose (SyRIP) dataset with small yet diverse real infant images as well as generated synthetic infant poses and (2) a multi-stage invariant representation learning strategy that could transfer the knowledge from the adjacent domains of adult poses and synthetic infant images into our fine-tuned domain-adapted infant pose (FiDIP) estimation model. In our ablation study, with identical network structure, models trained on SyRIP dataset show noticeable improvement over the ones trained on the only other public infant pose datasets. Integrated with pose estimation backbone networks with varying complexity, FiDIP performs consistently better than the fine-tuned versions of those models. One of our best infant pose estimation performers on the state-of-the-art DarkPose model shows mean average precision (mAP) of 93.6¹.

I. INTRODUCTION

Current efforts in machine learning, especially with the recent waves of deep learning models introduced in the last decade, have obliterated records for regression and classification tasks that have previously seen only incremental accuracy improvements. However, this performance comes at a large data cost. There are many other applications that would significantly benefit from the deep learning, where data collection or labeling is expensive and limited. In these domains, which we refer to as “Small Data” domains, the challenge we facing is how to learn efficiently with the same performance with less data. One example of these applications with the small data challenges is the problem of infant pose estimation. In infants, long-term monitoring of their poses provide information about their health condition and accurate recognition of these poses can lead to a better early developmental risk assessment and diagnosis [19], [5]. Both motor delays and atypical movements are presented in children with cerebral palsy and are risk indicators for autism spectrum disorders [30], [25].

¹The code is available at: github.com/ostadabbas/Infant-PoseEstimation. The SyRIP dataset can be downloaded at: Synthetic and Real Infant Pose (SyRIP).

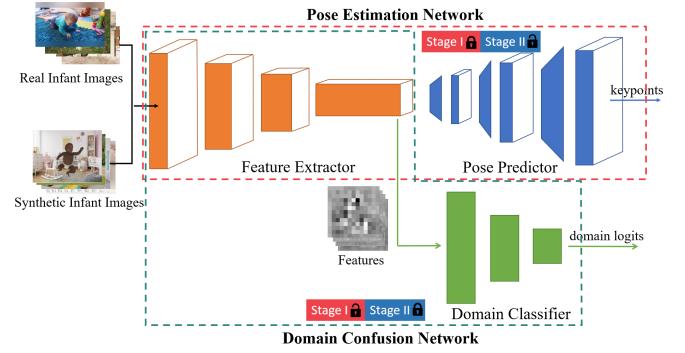


Fig. 1: An overview architecture of our fine-tuned domain-adapted infant pose (FiDIP) framework, composed of two sub-networks: pose estimation network (red-dot box) and domain confusion network (blue-dot box). Main components of FiDIP include a feature extractor (orange), a pose predictor (blue), and a domain classifier (green).

However, current publicly available human pose datasets are predominantly from scenes such as sports, TV shows, and other daily activities performed by adult humans, and none of these datasets provides any specific infants or young children pose images. Beside privacy issues which hamper large-scale data collection from infant, infant pose images differ from available adult pose datasets due to the notable differences in their pose distribution (see Fig. 5(a)) compared to the common adult poses collected from surveillance viewpoints [12]. These differences mainly come from (1) shorter limbs and completely different bone to muscle ratio compared to adults; (2) different activities, appearances, and environmental contexts, which together result in sub-optimal performance of the pre-trained models trained on adult poses when tested on infant images (see Section V) with either over-prediction or under-prediction of the limb sizes.

In this paper, towards building a robust infant pose estimation model, we propose a strategy to transfer the pose learning of the existing adult pose estimation models into the infant poses. It includes a hybrid synthetic and real infant pose dataset built based on our cross domain inspired augmentation (CDIA) approach and a fine-tuned domain-adapted infant pose (FiDIP) estimation network as shown in Fig. 1, which is a data-efficient inference model bootstrapped on both transfer learning and synthetic data augmentation approaches. In this paper, we address the critical small data problem of infant pose estimation by making the following contributions:

- Building and publicly releasing a novel full-annotated

hybrid synthetic and real infant pose (SyRIP) dataset via a proposed cross domain inspiration augmentation technique. SyRIP includes a diverse set of real and synthetic infant images, which benefits from (1) appearance and pose of real infants in images scrapped from web, and (2) the augmented variations in view points, poses, backgrounds, and appearances by synthesizing a set of infant avatars. SyRIP provides advantages over the existing (and very limited) infant pose datasets, by clearly improving the performance of the models trained on it.

- Proposing a fine-tuned domain-adapted infant pose (FiDIP) framework build upon a two-stage training paradigm. In the stage I of training, we fine-tune a pre-trained synthetic/real domain confusion network in a pose-unsupervised manner. In the stage II, we fine-tune a pre-trained pose estimation model under the guidance of stage I trained domain confusion network. Both networks are updated separately in iterative way.
- Achieving two invariant representation learning goals simultaneously. In the FiDIP network, there exist two transfer learning tasks: (1) from adult pose domain into the infant pose domain, and (2) from synthetic image domain into the real image domain. We fine-tune the pose estimation network by constraining that to extract features with common domain knowledge between synthetic and real data.
- Extensive experiments on the evaluation of each proposed component of FiDIP test on real infant pose images, which show that our method provides consistent performance improvement when applied on the existing state-of-the-art (SOTA) human pose estimation models with both complex and light-weight pose predictors. This allows the implementation of FiDIP on embedded systems (e.g. baby monitors) for long-term as well as real-time pose monitoring of infants.

II. RELATED WORK

a) Infant Pose Estimation.: For applications that require infant posture/motion analysis, the current approaches are dominantly based on (real-time or recorded) visual observation by the infant's pediatrician or the use of contact-based inertial sensors. Meanwhile, there exist very few recent attempts initiated by the computer vision community to automatically perform pose estimation and tracking on videos taken from infants. In [8], the authors estimated 3D body pose of infants in depth images for their motion analysis purpose, however they only evaluated their method on simple supine positions from limited number of subjects. They employed a pixel-wise body part classifier using random ferns to predict infant's 3D joints in order to automate the task of motion analysis for identifying infantile motor disorders. In [7], the authors presented a statistical learning method called 3D skinned multi-infant linear (SMIL) body model using incomplete low quality RGB-D sequence of freely moving infants. The specific dataset they used is provided in [6], where users mapped real infant movements to the SMIL

model with natural shapes and textures, and generated RGB and depth images with 2D and 3D joint positions. However, both of these works rely heavily on having access to the RGBD data sequence, which is difficult to obtain and hinder the use of these algorithms in regular webcam-based baby monitoring systems.

b) Synthetic Human Pose Data Generation.: Synthesizing complicated articulated 3D models such as a human body has been drawing huge attention lately due to its extensive applications in studying human poses, gestures, and activities. Among benefits of synthesizing data is the possibility to automatically generate enough labeled data for supervised learning purposes, especially in small data domains [23]. In [13], the authors introduce a semi-supervised data augmentation approach that can synthesize large-scale labeled pose datasets using 3D graphical engines based on a physically-valid low dimensional pose descriptor. As introduced in [20], 3D human poses can be reconstructed by learning a geometry-aware body representation from multi-view images without annotations. Another research trend in synthesizing human pose images is simulating human figures by employing generative adversarial network (GAN) techniques. The authors in [15] present a two-stage pose-guided person generation network to integrate pose by feeding a reference image and a novel pose into a U-Net-like network to generate a coarse reposed person image, and refine image by training the U-Net-like generator in an adversarial way. In these works, however, neither the generated human avatars nor the reconstructed poses are able to accurately adapt to the infant style. Additionally, these GAN-based approaches of synthetic human figures do not have the capabilities of simulating complicated poses regularly taken by infants.

Based on the above-mentioned challenges in achieving a robust infant pose estimation model and the shortcomings of the prior arts, we address the problem by contributing: (1) a hybrid real and synthetic infant pose dataset (SyRIP) that benefits from both realistic poses/appearances as well as synthesizing augmentation, (2) a fine-tuned domain-adapted infant pose (FiDIP) approach which shows consistent improvement over conventional fine-tuning evaluated on several SOTA backbones (see Fig. 2).

III. SYRIP: SYNTHETIC/REAL INFANT POSE DATASET FOR POSE DATA AUGMENTATION

As stated earlier, there is a shortage of labeled infant pose dataset, and despite recent efforts in developing them, a versatile dataset with different and complex poses to train a deep network on is yet to be built. The only publicly-available infant image dataset is MINI-RGBD dataset [6], which provides only 12 synthetic infant models with continuous pose sequences. However, beside having simple poses, MINI-RGBD sequential feature leads to a small variation in the poses between adjacent frames and the poses of whole dataset are mainly repeated. In Fig. 3(a), we show the distribution of body poses of MINI-RGBD dataset and observe that poses in this dataset are simple and lack variations. Both its simplicity and being exclusively synthetic would cause

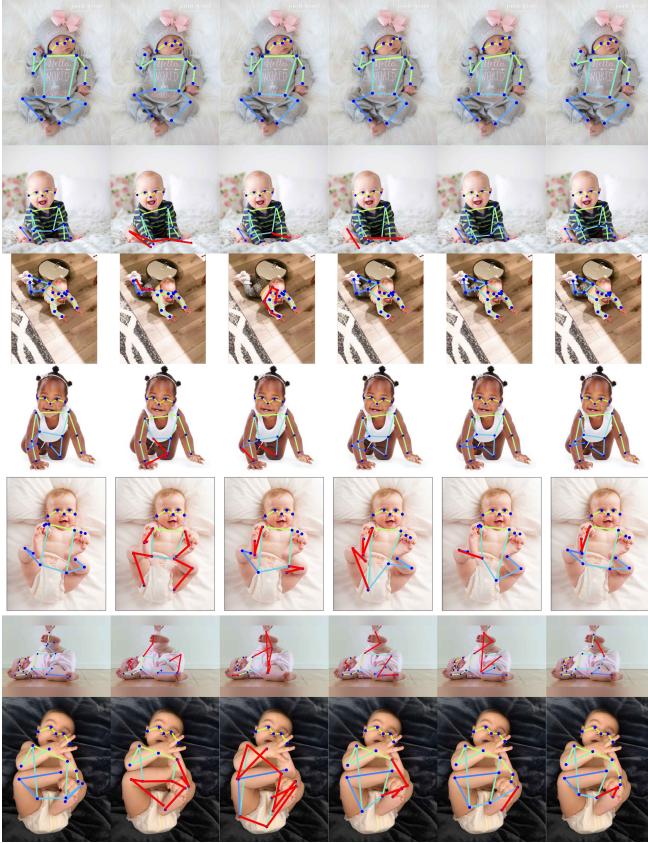


Fig. 2: Samples of infant pose prediction results of DarkPose: AP=65.9 (2nd column), FasterR-CNN: AP=70.1 (3rd column), SimpleBaseline: AP=82.4 (4rd column), DarkPose: AP=88.5 (5th column), and our SimpleBaseline+FiDIP: AP=91.1 (6th column) on SyRIP Test100, which are listed in Table II. The 1st column is the visualization of groundtruth. Incorrect predictions are highlighted in red. Note: more visualized results are given in *Supplementary Materials*.

the pose estimation models trained on MINI-RGBD to not generalize well to the real-world infant images.

However, collecting and fully annotating a large-scale real infant pose dataset comparable to the size of adult pose datasets, such as Microsoft COCO [11], MPII [1], LSP [10], and FLIC [22], is very challenging, due to the privacy concerns that have also limited the number of samples in the web. To address this data limitation, we present a hybrid dataset forming strategy by mixing both real and synthetic infant pose data to form our SyRIP dataset, which exploits not only the real infant pose and appearance information, but also the data augmentation flexibility through using a set of synthetic infant models.

A. Real Infant Pose Data Gathering

Due to the difficulties in controlling infant movements as well as critical privacy concerns in collecting images from someone's child, access to infant images with wide variety of poses is limited. For real portion of the SyRIP dataset, we look for publicly available yet scattered real infant images from sources such as *YouTube* and *Google Images*. The biggest benefit of this collection method is that the diversity of the infant poses is guaranteed to the greatest extent. We choose infant (newborn to one year old) in various poses and

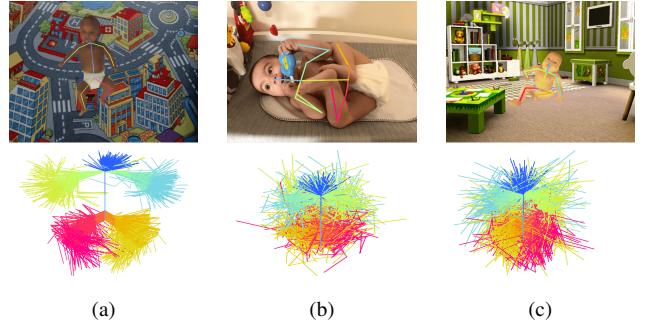


Fig. 3: A sample image and pose distribution of (a) MINI-RGBD dataset, (b) the real part of the SyRIP dataset, and (c) the synthetic part of the SyRIP dataset. The first row shows a sample image from each dataset with its groundtruth labels. The second row shows the pose distribution of 200 images that are randomly selected from each dataset, in which colors of different body parts correspond to the colors of body parts of figures in the first row. For the pose distribution, we normalize all images based on the infant bounding box to scale them into similar sizes, then align them based on their torso with upward head. To better represent the poses, we also ignore the points for ears and eyes when visualizing the joints.

many different backgrounds.

We manually query YouTube and download more than 40 videos with different infants, and then split each video sequence to pick about 12 frames containing different poses. Finally, about 500 images including more than 50 infants with different poses from those frames are collected. We also select about 200 high-resolution images containing more than 90 infants from the Google Images. Compared to images taken from the YouTube videos, images from Google Images with higher resolution can be used to improve the quality of the whole dataset. The pose distribution of the real part of the SyRIP dataset is shown in the Fig. 3(b). Obviously, these poses are more diverse than those in the MINI-RGBD dataset.

B. Cross Domain Inspired Synthetic Data Augmentation

200 real images is far too small to train a deep neural network and even not enough to fine-tune a pose estimation model with deep structure. Simulation seems to be a valid way to augment the dataset [24], however it comes out to be challenging for infants as there are neither many infant 3D scans available to augment their appearance data, nor any infant motion captured movements to augment their pose data. Here we propose a cross domain inspired synthetic augmentation approach for infant pose data simulation. The pipeline of synthetic augmentation is illustrated in Fig. 4.

We employ the SMIL model [6] for our synthetic data generation, which has $N = 6890$ vertices and $K = 23$ joints, and can be parameterized by the pose coefficients $\theta \in \mathbb{R}^{3(K+1)}$, where $K+1$ stands for body joints and one more joint (i.e. pelvis, is the root of the kinematic tree) for global rotation, and the shape coefficients $\beta \in \mathbb{R}^{20}$, representing the proportions of the individual's height, length, fat, thin, and head-to-body ratio. The infant mesh is then given as $M(\beta, \theta)$ and a synthetic image I_{syn} is generated through the imaging process \mathcal{I} as:

$$I_{syn} = \mathcal{I}(M(\beta, \theta), C(d, f), Tx, Bg), \quad (1)$$

where C stands for the camera parameters depending on the camera principal point d and focal length f . Tx stands for the texture and Bg stands for the background. We augment the camera parameter with random position with a fixed focal length. For the background, we pick 600 scenarios approximately related to infant indoor and outdoor activities from LSUN dataset [28]. Unfortunately, SMIL provides only limited appearances and simple pose parameters. There are no known infant motion capture data neither extra infant appearances for the SMIL model. To augment these parameters, we employ references from the neighboring domains as following:

Inspiration from the real infant pose domain: Although we do not have a versatile 3D infant pose data for pose augmentation, however these infant specific poses are actually reflected in the real infant images scrapped from web though in 2D. We first employ the SMPLify-x approach [18] to lift these 2D poses into the SMIL pose θ by minimizing the cost function as:

$$L = L_J(\beta, \theta; C, j_{2D}) + \lambda_\theta L_\theta(\theta) + \lambda_\beta L_\beta(\beta) + \lambda_\alpha L_\alpha(\theta), \quad (2)$$

where C is intrinsic camera parameters, λ_θ , λ_β , and λ_α are weights for specific loss terms, as described in [18]. Eq. (2) is the sum of four loss terms: (1) L_J a joint-based data term, which is the distance between groundtruth 2D joints j_{2D} and the 2D projection of the corresponding posed 3D joints of SMIL for each joint, (2) L_θ defined as a mixture of Gaussians pose prior learnt from 37,000 adult poses [18], (3) a shape penalty L_β , which is the Mahalanobis distance between the shape prior of SMIL and the shape parameters being optimized, and (4) a pose prior penalizing elbows and knees L_α . Therefore, we can augment the synthetic infant pose and shape via learned parameter from the real images.

Inspiration from the adult pose domain: Even different in size, infants still share the similar kinematic structure as adults and most adult poses are kinematically compatible with infants. With the same topology as the adult template SMPL [14], many existing adult scans' textures can be transferred directly into the infant models. A valid concern could be that with adult's textures the model will be no infant-like. However, we argue that with limited data, these borrowed variations can prevent overfitting and improve the model robustness. Therefore, we not only utilize the 12 infant textures (naked only with diaper) provided by MINI-RGBD dataset, but also augment appearance with adult textures from 478 male and 452 female clothing images coming from synthetic humans for real (SURREAL) dataset [24].

During synthesizing, we manually filter out the unnatural/invalid generated infant bodies and add random noise term into augmented pose data to further increase its variance. We visualize the pose distribution of the SyRIP synthetic subset, in Fig. 3(c), to make sure that the poses in our synthetic dataset has enough variations.

Finally, a new infant pose dataset, synthetic and real infant pose (SyRIP), is built up including both real and synthetic

images that display infants in various positions, and utilize it to train pose estimation models with our proposed FiDIP method. Our process includes a training part consists of 200 real and 1000 synthetic infant images, and a test part with 500 real infant images, all with fully annotated 2D body joints. Infants in these images have many different poses, like crawling, lying, sitting, and so on. The pose distributions of MINI-RGBD, SyRIP real part and SyRIP synthetic part are shown in Fig. 3, in which SyRIP dataset shows noticeably more pose variations compared the existing infant pose dataset, MINI-RGBD.

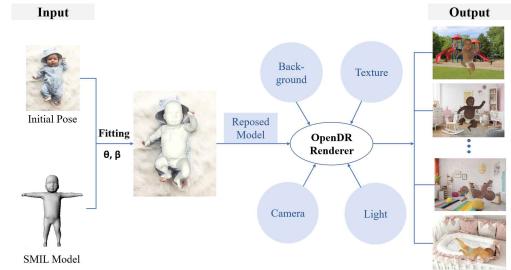


Fig. 4: Our pipeline of synthetic infant image generation. 3D infant body models are posed by fitting SMIL model pose and shape parameters into real infant images. The generated model is reposed by adding variances to pose coefficients, θ . Then output images are rendered using random background images, texture maps on the body, lighting, and camera positions.

IV. FiDIP: FINE-TUNED DOMAIN-ADAPTED INFANT POSE ESTIMATION

Our FiDIP approach makes use of an initial pose estimation model trained on the abundant adult pose data, then fine-tunes that model on the augmented dataset, which consists of a small amount of real infant pose data and a series of pose-diverse synthetic infant images. For the augmented dataset, a domain adaptation method is proposed to align features of synthetic infant data with the real-world infant images. As the number of images in our dataset is limited, we only update a few layers of that network to fine-tune that for infant pose estimation rather than re-training the whole adult pose estimation network.

A. FiDIP Framework

Our FiDIP framework is integrable with any existing encoder-decoder pose model. As illustrated in Fig. 1, a pose estimation model with feature extractor as its encoder and pose estimator as its decoder could apply FiDIP by introducing a domain classification head. Then the entire model can be treated as two sub-network: pose estimation network and domain confusion network. Pose estimation network could be any pose model including SimpleBaseline [27], DarkPose [29], and Hourglass [17]. The domain confusion network, which is composed of a feature extractor shared with the pose estimation component and a domain classifier, is added to enforce the images in the real or synthetic domain being mapped into a same feature space after feature extraction. Domain classifier is designed to be a binary classifier with only three fully connected layers to distinguish whether the input feature belongs to a real or

synthetic image. In particular, the domain confusion network assists pose estimation network during training. At test time, only the pure pose model (pose estimation network) works independently.

B. Network Training

The FiDIP training procedure consists of initialization session and a formal training session where the domain classifier and feature extractor are trained in a circular way.

Model initialization: The pose estimation component of FiDIP network is already pre-trained on adult pose images from COCO dataset [11]. Since our training strategy is based on the use of fine-tuning as a means for transfer learning, to avoid unbalanced components’ updating during fine-tuning, the domain classifier part of our domain confusion sub-network also needs to be pre-trained on both real and synthetic data from adult humans in advance. This combination dataset includes real adult images from the validation part of COCO dataset and some part of SURREAL dataset [24]. During this pre-training, the feature extractor part stays frozen, and only the weights for domain classifier will be initialized. The following stages are done after this initialization.

Formal training session: In this session, for each iteration the network is updated in a circular way with two stages:

Stage I. In this stage, we lock pose estimation sub-network and fine-tune the domain classifier of domain confusion sub-network based on the current performance of feature extractor using infant real and synthetic pose data. The objective of this stage is to obtain a domain classifier for predicting whether the features are from a synthetic infant image or real one. Since the pose estimation network is locked and only domain classifier is to be optimized, the optimization objective in this stage is the loss of domain classifier L_D , which is calculated by the binary cross entropy:

$$L_D = -\frac{1}{N} \sum_{i=1}^N d_i \cdot \log f(s_i) + (1 - d_i) \cdot \log(1 - f(s_i)), \quad (3)$$

where s_i is the score of i th feature belonging to synthetic domain, d_i is the corresponding groundtruth, $f(\cdot)$ represents the Sigmoid function, and N is the batch size.

Stage II. The pose estimation network is to be fine-tuned with locked domain classifier in this stage. We try to refine the feature extractor to not only affect the pose predictor but also confuse the domain classifier. We leverage the domain classifier updated at stage I to promote the feature extractor to retain the ability to extract keypoints’ information during the fine-tuning process, but also to ignore the differences between the real domain and the synthetic domain. An adversarial training method, which is proposed in [4], is utilized to pushing features from synthetic images and real images into a common domain. A gradient reversal layer (GRL) is introduced to minimize the pose loss (L_P).

To train our domain classifier in a balance way, we propose a balancing strategy by increasing the weight of real data during training. The L_P loss, which measures the mean

squared error between predicted heatmap/coordinates \hat{y}_i and targeted heatmap/coordinates y_i for each keypoint i , is:

$$L_P = \frac{1}{N} \sum_{i=1}^N S(I_i)(\hat{y}_i - y_i)^2, \quad (4)$$

where $S(I_i)$ is the scaling factor in the domain indicator I_i . It simultaneously maximizes the domain loss (L_D), so that the features representing both synthetic and real domains become similar. The optimization objective is:

$$L(\theta_f, \theta_y, \theta_d) = L_P(\theta_f, \theta_y) - \lambda L_D(\theta_f, \theta_d), \quad (5)$$

where λ controls the trade-off between the two losses that shape the features during fine-tuning. θ_f , θ_y , and θ_d represent parameters of feature extractor, pose predictor, and domain classifier, respectively.

V. EXPERIMENTAL EVALUATION

Our solution towards a robust infant pose estimation comes from two main contributions: (1) SyRIP dataset and (2) FiDIP approach. In this section, we evaluate each component specifically.

A. Datasets

Two infant datasets are employed in our evaluation, one is the only public infant pose dataset, called MINI-RGBD, and the other one is our SyRIP dataset. We also employ 1904 samples from COCO val2017 [11] and 2000 random images from SURREAL [24] during our pre-training stage.

MINI-RGBD dataset [6] has 12 synthetic infant models with their continuous pose sequences. SyRIP dataset includes 700 real infant images with representative poses via manually selection and 1000 synthesized infants. For a reliable evaluation, we keep a large portion of 500 real infant data as test set which we call Test500 for common our common test. The rest 200 real with the synthetic infant data is used at the training set. One observation in our study is that, many pre-trained models are able to accurately estimate infant poses that are similar to adult poses, however not for infant-unique poses such as bend legs over the chest. To evaluate a model’s performance over specific infant poses, we collect a challenging subset with 100 complex yet typical infant poses from Test500 which we call Test100. Some samples are provided in *Supplementary Materials*.

It is clear that the number of images in our test set is much smaller compared to the datasets used in other human pose estimation studies. Indeed, due to the aforementioned limitations caused by privacy, security, and other objective conditions, obtaining sufficient amount of infant pose images (that can publicly get access to) is an ongoing challenge, which makes our application a clear example in “Small Data” domain. We make up for the lack of data scale by enriching the poses, characters, and scenes in our SyRIP dataset.

B. Implementation Details

We employ several SOTA pose estimation structures with varying complexity as our backbone network, including the SimpleBaseline-50 [27], DarkPose [29] and MobileNetV2

TABLE I: Performance comparison of three SOTA pose estimation models (SimpleBaseline, DarkPose, Pose-MobileNet) fine-tuned on MINI-RGBD, SyRIP-syn (synthesized data only) and SyRIP whole set and tested on SyRIP Test100.

Train Set	Method	Backbone	Input size	AP	AP50	AP75	AR	AR50	AR75
MINI-RGBD	SimpleBaseline	ResNet-50	384x288	69.2	95.8	78.0	72.4	97.0	81.0
SyRIP-syn				90.1	98.5	97.2	91.6	99.0	98.0
SyRIP				91.1	98.5	98.5	92.6	99.0	99.0
MINI-RGBD	DarKPose	HRNet-W48	384x288	85.2	98.6	95.3	87.0	99.0	96.0
SyRIP-syn				91.4	98.5	98.5	92.7	99.0	99.0
SyRIP				92.7	98.5	98.5	93.9	99.0	99.0
MINI-RGBD	Pose-MobileNet	MobileNetV2	224x224	12.3	38.1	3.8	21.6	52.0	14.0
SyRIP-syn				60.3	91.1	62.7	68.4	95.0	72.0
SyRIP				78.9	97.2	90.6	84.2	98.0	94.0

[21] to reflect the general effect of our FiDIP framework. We add our domain classifier which has 3 fully connected layers on top of the backbone output features. For DarkPose, we choose the highest resolution branch. During training, we employ Adam optimizer with learning rate of 0.001. The batch size and epoch for initialization session are 128 and 1, respectively. While, for formal training session, there are 100 epochs and 64 images in a batch. During the Stage II, we set GRL parameter λ as 0.0005, and freeze the first three layers (Res1, Res2, and Res3) of the feature extractor in our detailed ablation study. As for evaluation metric, we employ mean average precision (mAP) [11] over 10 thresholds of the object keypoint similarity (OKS), which is the distance between predicted keypoints and ground truth keypoints normalized by the scale of the person.

C. Evaluation Over SyRIP

We gauge the SyRIP quality by specifically evaluating the effect of its synthetic data as well as its real and synthetic hybrid data. In a straight forward way, we compare identical models fine-tuned on SyRIP or MINI-RGBD datasets to compare their performances as shown in Table I.

To evaluate SyRIP quality, we employ three SOTA pose estimation models to fine-tune on MINI-RGBD, SyRIP-syn (synthetic portion only) and SyRIP whole set and compare their performance as shown in Table I. From the result, we can see that with limited synthesized appearances and limited poses, the model tuned on MINI-RGBD is easily overfitted with even lower performance than the original model. In comparison, in our CDIA approach by extensively learning from neighboring domains, the data variation is increased and even with our synthetic infant data alone, 'SyRIP-syn' and without any adaptation, the model performance is still improved. Additional real infant data as in full SyRIP set, further increase the performance that indicates the benefit of our hybrid strategy. All these improvements are observed on all tested models with varying computational complexities.

D. Evaluation over FiDIP

For the infant pose estimation problem, two hypotheses may be assumed: (1) 2D human pose estimation models trained on the large-scale public datasets will be universally effective on different subjects, including infants. (2) If not, they can be fine-tuned with a few samples from the target domain to achieve high performance. In this section, we evaluate these hypotheses by comparing: (a) FiDIP with

the SOTA pre-trained models; (b) FiDIP ablation study; (c) FiDIP with the conventional fine-tuning approach. For fair comparison, all models are trained on SyRIP if needed and the performance advantage purely comes from the approaches.

a) *Comparison with the SOTA general purpose pose estimation models:* We compare our FiDIP model with a ResNet50 backbone [27] with pre-trained SOTA approaches as shown in Table II. Obviously, most models are well-performed on SyRIP Test500, which indicates infant and adult share many common poses. However for infant-specific poses in Test100, their performance drops dramatically as these poses are rarely seen among adults. In comparison, our approach (FiDIP) shows noticeably better results in both Test100 and Test500. We can see that pre-trained SOTA human pose models are not universally effective and infant pose estimation can be improved significantly via our approach.

We also provide qualitative visualizations of our Simple-Baseline+FiDIP model on SyRIP test dataset compared to the Faster R-CNN, DarkPose, and SimpleBaseline models performance in Fig. 2. Simple poses, such as the examples in the 1st row of Fig. 2, are predicted accurately by almost all of the SOTA models. However, in infant's daily activities, their poses are often varied and more complex, especially in their lower body. DarkPose model based on ResNet-50 with 128×96 input size (2nd column) and Faster R-CNN model based on ResNet-50 (3rd column) trained on the adult datasets, show obvious inaccuracies in localizing the position of infant's legs and feet. Even SimpleBaseline and DarkPose based on HRNet [2] models with 384×288 input size are unable to keep high performance of infant lower body estimation. SimpleBaseline+FiDIP has much greater chance of inferring keypoints correctly for infant pose images than other models as shown in Fig. 2.

b) *Ablation study:* Table III investigates the performance of alternative choices in the FiDIP on SimpleBaseline-50 model, where method **n** is our well-performed FiDIP model as reported in Table II.

Domain Adaptation. We explore whether the domain adaptation method we implement can effectively overcome the difference between feature spaces of the real (R) domain and synthetic (S) domain in our SyRIP training dataset, so we test on 700 real images and 1000 synthetic images from the whole SyRIP dataset(1200 training + 500 testing) for easier observation. Methods that contain domain adaptation

TABLE II: Performance comparison between SimpleBaseline model applied FiDIP method and the SOTA pose estimators on the COCO Val2017 and SyRIP test datasets.

Pose Estimation Model	Backbone Network	Input Image Size	COCO_Val2017	SyRIP_Test500	SyRIP_Test100					
			AP	AP	AP	AP50	AP75	AR	AR50	AR75
Faster R-CNN [26]	ResNet-50-FPN	Flexible	65.5	93.4	70.1	97.7	73.8	-	-	-
Faster R-CNN [26]	ResNet-101-FPN	Flexible	66.1	91.9	64.4	95.2	71.5	-	-	-
DarkPose [29]	ResNet-50	128×96	62.6	95.2	65.9	94.8	66.7	69.2	96.0	71.0
DarkPose [29]	HRNet-W48	128×96	71.9	97.4	82.1	98.6	92.2	83.6	99.0	93.0
DarkPose [29]	HRNet-W32	256×192	75.6	97.7	88.5	98.4	98.4	90.1	99.0	99.0
DarkPose [29]	HRNet-W48	384×288	76.9	98.0	88.5	98.5	98.5	90.0	99.0	99.0
SimpleBaseline [27]	ResNet-50	256×192	70.4	97.3	80.4	98.5	92.2	82.5	99.0	94.0
SimpleBaseline [27]	ResNet-50	384×288	72.2	97.6	82.4	98.9	92.2	83.8	99.0	93.0
RMPE [3]	VGG_SSD	500×500	61.8	76.2	76.3	82.4	78.3	-	-	-
UDP [9]	HRNet-W32	256×192	75.2	81.2	79.8	86.2	88.4	71.3	73.2	74.0
UDP [9]	ResNet-50	256×192	71.7	83.4	78.2	80.2	77.4	75.1	75.4	76.7
UDP [9]	ResNet-152	384×288	74.7	84.2	79.1	81.5	82.8	81.1	80.4	79.6
SimpleBaseline + FiDIP (Ours)	ResNet-50	384×288	59.6	98.3	91.1	98.5	98.5	92.6	99.0	99.0

Best results are highlighted in bold fonts.

TABLE III: Ablation study of FiDIP on SyRIP Test100 dataset with resolution of 384×288. DC stands for domain classifier. SimpleBaseline-50 [27] is provided here for a baseline comparison.

Method	Training Data	Domain Adaptation	Pre-train DC	Update Layers	SyRIP Test-AP
SB-50(*)	-	-	-	-	82.4
a	1000 Syn	×	-	Res 4, 5	84.1
b	1000 Syn	×	-	Res 5	85.3
c	1000 Syn	✓	✗	Res 4, 5	84.6
d	1000 Syn	✓	✓	Res 4, 5	85.3
e	1000 Syn	✓	✗	Res 5	86.3
f	1000 Syn	✓	✓	Res 5	85.5
g	200 Real	✗	-	Res 4, 5	87.1
h	200 Real	✗	-	Res 5	86.9
i	1200 R+S	✗	-	Res 4, 5	90.1
j	1200 R+S	✗	-	Res 5	90.0
k	1200 R+S	✓	✗	Res 5	90.2
l	1200 R+S	✓	✓	Res 5	90.3
m	1200 R+S	✓	✗	Res 4, 5	90.3
n	1200 R+S	✓	✓	Res 4, 5	91.1

* SimpleBaseline-50

show higher AP than other method without domain adaptation. t-SNE [16] is used to visualize the distributions of extracted features for original SimpleBaseline-50, method **i**, and method **n** in Fig. 5. Obviously, the FiDIP method embedded with domain adaptation component can align the feature distribution better than other networks.

Update Layers. Freezing weights of the first few layers of the pre-trained network is a common practice when fine-tuning network with an insufficient amount of training data. The first few layers are responsible to capture universal features like curves and edges, so we fix them to enforce our network to focus on learning dataset-specific features in the subsequent layers at Stage II. We explore the effect of updating different numbers of last few layers of network on the performance of the trained model. In Table III, for method **m** and **n**, the ResNet 4th and 5th blocks of our feature extractor (ResNet-50) are updated, while the first four ResNet blocks are fixed and only the weights of last one block are updated in method **k** and **l**. We observe that method **m**, **n** perform much better than the other two.

c) Comparison with Direct Fine-Tuning: A classical approach for transfer learning is a straight forward fine-tuning. Here, we employ three SOTA backbones for our pose estimation models with varying complexity, MobileNetV2,

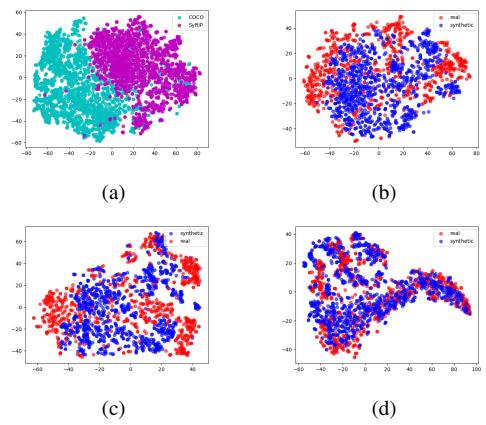


Fig. 5: t-SNE visualized features of (a) 700 random samples of COCO Val2017 vs. 700 real images of SyRIP, and 1000 synthetic images vs. 700 real images of SyRIP dataset extracted by (b) original SimpleBaseline-50, (c) method **i** (fine-tuning without domain adaptation), and (d) method **n** (fine-tuning with domain adaptation).

DarkPose, and SimpleBaseline-50 and compare the FiDIP version and the fine-tuned version head to head with result shown in Table IV. To achieve pose estimation goal on backbone MobileNetV2, the Pose-MobileNet is built by adding a pose regressor as a decoder behind MobileNetV2. We initially train it on COCO Train2017 to get a pre-trained model and then fine-tune or apply FiDIP method to Pose-MobileNet on SyRIP dataset.

VI. CONCLUSION

In this paper, we present a solution towards robust infant pose estimation, which includes an infant dataset SyRIP with hybrid synthetic and real data and a FiDIP strategy to transfer learn from existing adult models and datasets. Our FiDIP framework consists of a pose estimation sub-network to leverage transfer learning from a pre-trained adult pose estimation network and a domain confusion sub-network for adapting the model to both real infant and synthetic infant datasets. With identical network structure, we compared their performance when trained on our SyRIP dataset and on the only other publicly available infant pose dataset MINI-RGBD respectively to show the benefit of dataset forming strategy with high data scarcity challenge. With identical

TABLE IV: Evaluating the generality of our FiDIP method to different SOTA models on the SyRIP Test100.

Method	Backbone	Input size	# Params	GFLOPs	AP	AP50	AP75	AR	AR50	AR75
SimpleBaseline	ResNet-50	384x288	32.42M	20.23	82.4	98.9	92.2	83.8	99.0	93.0
SimpleBaseline + Finetune					90.1	98.5	97.2	91.6	99.0	98.0
SimpleBaseline + FiDIP					91.1	98.5	98.5	92.6	99.0	99.0
DarkPose	HRNet-W48	384x288	60.65M	32.88	88.5	98.5	98.5	90.0	99.0	99.0
DarkPose + Finetune					92.7	98.5	98.5	93.9	99.0	99.0
DarkPose + FiDIP					93.6	98.5	98.5	94.6	99.0	99.0
Pose-MobileNet	MobileNetV2	224x224	3.91M	0.46	46.5	85.7	45.6	56.2	89.0	59.0
Pose-MobileNet + Finetune					78.9	97.2	90.6	84.2	98.0	94.0
Pose-MobileNet + FiDIP					79.3	99.0	89.4	84.1	99.0	92.0

dataset, we compared FiDIP approach with the fine-tuning approach to show the advantages of FiDIP across multiple pose estimation models with various complexities.

REFERENCES

- [1] M. Andriluka, L. Pishchulin, P. Gehler, and B. Schiele. 2d human pose estimation: New benchmark and state of the art analysis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3686–3693, 2014.
- [2] B. Cheng, B. Xiao, J. Wang, H. Shi, T. S. Huang, and L. Zhang. Bottom-up higher-resolution networks for multi-person pose estimation. *CoRR*, abs/1908.10357, 2019.
- [3] H.-S. Fang, S. Xie, Y.-W. Tai, and C. Lu. RMPE: Regional multi-person pose estimation. In *ICCV*, 2017.
- [4] Y. Ganin and V. Lempitsky. Unsupervised domain adaptation by backpropagation. In *International conference on machine learning*, pages 1180–1189. PMLR, 2015.
- [5] M. Hadders-Algra, A. W. K. Van den Nieuwendijk, A. Maitijn, and L. A. van Eykern. Assessment of general movements: towards a better understanding of a sensitive method to evaluate brain function in young infants. *Developmental Medicine & Child Neurology*, 39(2):88–98, 1997.
- [6] N. Hesse, C. Bodensteiner, M. Arens, U. G. Hofmann, R. 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)*, pages 0–0, 2018.
- [7] N. Hesse, S. Pujades, J. Romero, M. J. Black, C. Bodensteiner, M. Arens, U. G. Hofmann, U. Tacke, M. Hadders-Algra, R. Weinberger, et al. Learning an infant body model from rgb-d data for accurate full body motion analysis. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 792–800. Springer, 2018.
- [8] N. Hesse, A. S. Schröder, W. Müller-Felber, C. Bodensteiner, M. Arens, and U. G. Hofmann. Body pose estimation in depth images for infant motion analysis. In *2017 39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 1909–1912. IEEE, 2017.
- [9] J. Huang, Z. Zhu, F. Guo, and G. Huang. The devil is in the details: Delving into unbiased data processing for human pose estimation. In *The IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.
- [10] S. Johnson and M. Everingham. Clustered pose and nonlinear appearance models for human pose estimation. In *Proceedings of the British Machine Vision Conference*, 2010. doi:10.5244/C.24.12.
- [11] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014.
- [12] S. Liu and S. Ostadabbas. A vision-based system for in-bed posture tracking. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 1373–1382, 2017.
- [13] S. Liu and S. Ostadabbas. A semi-supervised data augmentation approach using 3d graphical engines. In *European Conference on Computer Vision*, pages 395–408, 2018.
- [14] M. Loper, N. Mahmood, J. Romero, G. Pons-Moll, and M. J. Black. Smpl: A skinned multi-person linear model. *ACM transactions on graphics (TOG)*, 34(6):1–16, 2015.
- [15] L. Ma, X. Jia, Q. Sun, B. Schiele, T. Tuytelaars, and L. Van Gool. Pose guided person image generation. In *Advances in neural information processing systems*, pages 406–416, 2017.
- [16] L. v. d. Maaten and G. Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(Nov):2579–2605, 2008.
- [17] A. Newell, K. Yang, and J. Deng. Stacked hourglass networks for human pose estimation. In *European conference on computer vision*, pages 483–499. Springer, 2016.
- [18] G. Pavlakos, V. Choutas, N. Ghorbani, T. Bolkart, A. A. A. Osman, D. Tzionas, and M. J. Black. Expressive body capture: 3d hands, face, and body from a single image, 2019.
- [19] H. F. Prechtel. Qualitative changes of spontaneous movements in fetus and preterm infant are a marker of neurological dysfunction. *Early human development*, 1990.
- [20] H. Rhodin, M. Salzmann, and P. Fua. Unsupervised geometry-aware representation for 3d human pose estimation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 750–767, 2018.
- [21] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4510–4520, 2018.
- [22] B. Sapp and B. Taskar. Modec: Multimodal decomposable models for human pose estimation. In *In Proc. CVPR*, 2013.
- [23] H. Su, C. R. Qi, Y. Li, and L. J. Guibas. Render for cnn: Viewpoint estimation in images using cnns trained with rendered 3d model views. *Proceedings of the IEEE International Conference on Computer Vision*, pages 2686–2694, 2015.
- [24] G. Varol, J. Romero, X. Martin, N. Mahmood, M. J. Black, I. Laptev, and C. Schmid. Learning from synthetic humans. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 109–117, 2017.
- [25] K. Vyas, R. Ma, B. Rezaei, S. Liu, M. Neubauer, T. Ploetz, R. Oberleitner, and S. Ostadabbas. Recognition of atypical behavior in autism diagnosis from video using pose estimation over time. In *2019 IEEE 29th International Workshop on Machine Learning for Signal Processing (MLSP)*, pages 1–6. IEEE, 2019.
- [26] Y. Wu, A. Kirillov, F. Massa, W.-Y. Lo, and R. Girshick. Detectron2. <https://github.com/facebookresearch/detectron2>, 2019.
- [27] B. Xiao, H. Wu, and Y. Wei. Simple baselines for human pose estimation and tracking. In *Proceedings of the European conference on computer vision (ECCV)*, pages 466–481, 2018.
- [28] F. Yu, Y. Zhang, S. Song, A. Seff, and J. Xiao. Lsun: Construction of a large-scale image dataset using deep learning with humans in the loop. *arXiv preprint:1506.03365*, 2015.
- [29] F. Zhang, X. Zhu, H. Dai, M. Ye, and C. Zhu. Distribution-aware coordinate representation for human pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7093–7102, 2020.
- [30] L. Zwaigenbaum, S. Bryson, and N. Garon. Early identification of autism spectrum disorders. *Behavioural brain research*, 251:133–146, 2013.

A. SUPPLEMENTARY MATERIALS

A. Complex Poses

Most of the infant poses are very different from those of adults. Especially because of the baby's softer body, the folded poses and occluded joints are more difficult to be recognized or predicted. Some of these typical poses selected from our SyRIP Test100 (complex poses collection) are shown in Fig. S1.

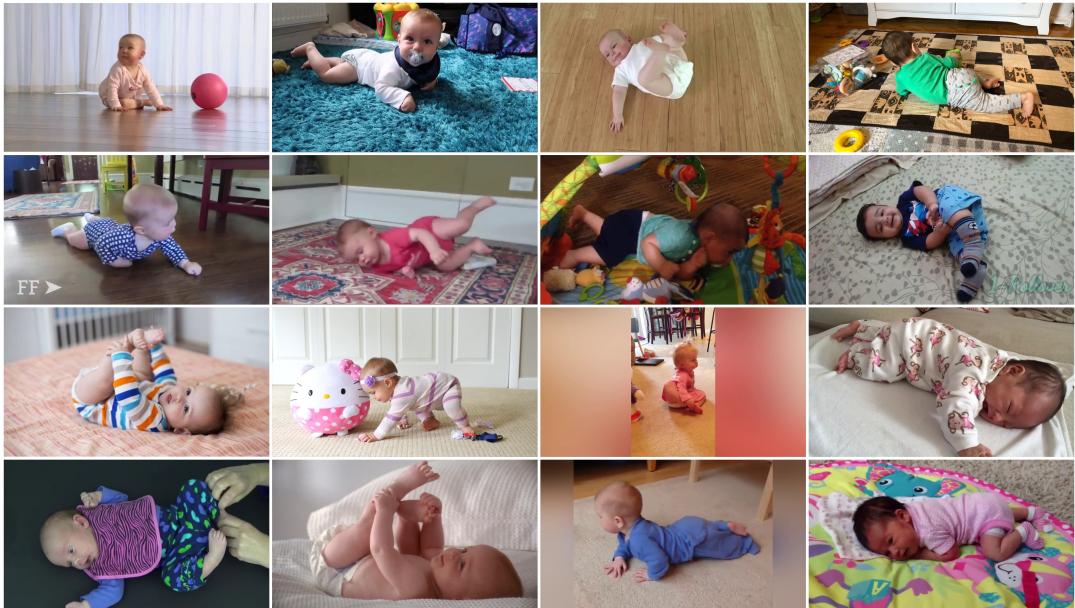


Fig. S1: Complex poses in SyRIP Test100.

B. Qualitative Results

In Fig. S2, we exhibit more visualized results for our SimpleBaseline+FiDIP model compared with the other well-performed pose estimation models (where their AP is higher than 90.0 on SyRIP Test500) listed in Table II.

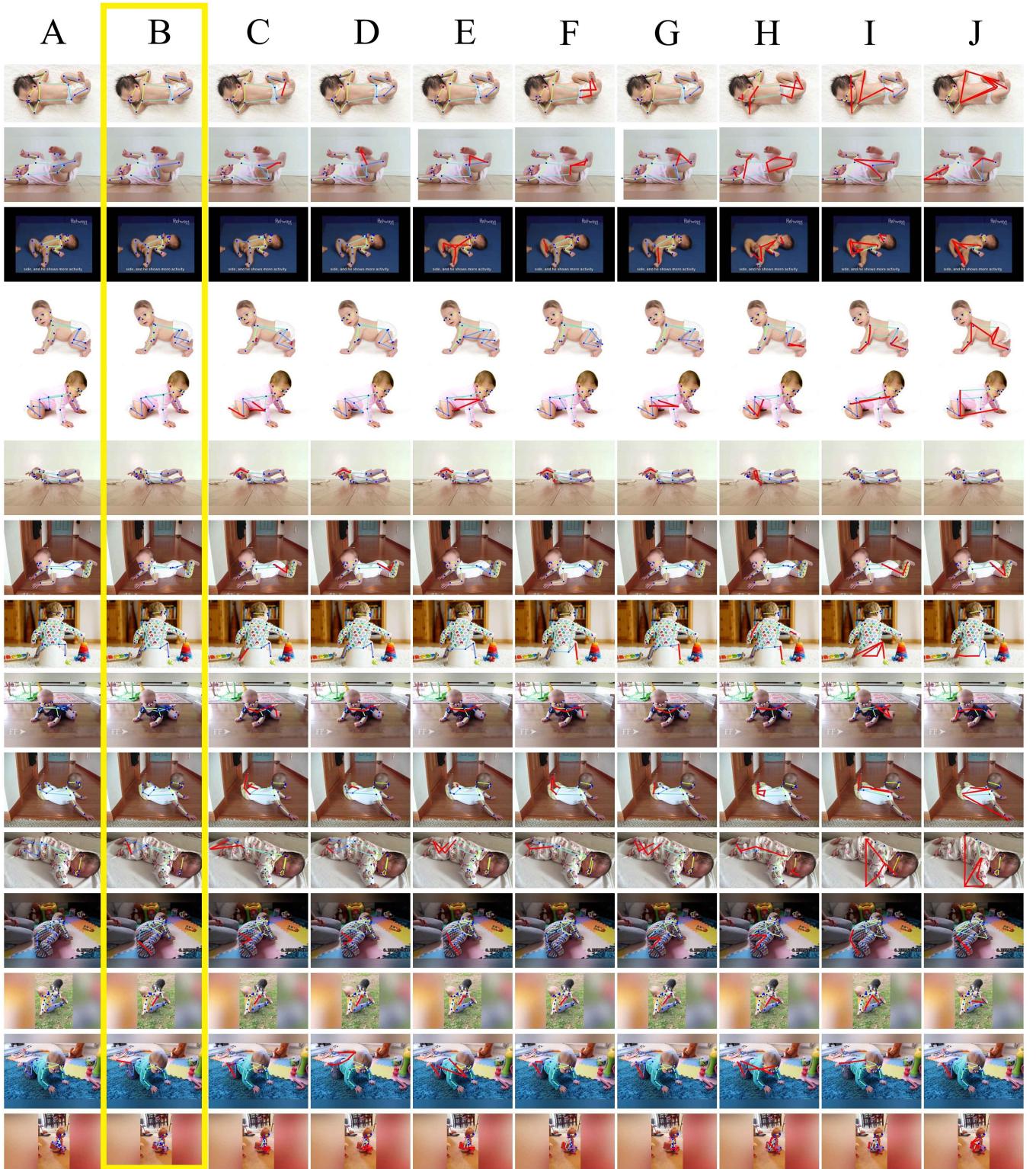


Fig. S2: Samples of infant pose prediction results of SOTA models on SyRIP Test500, which are listed in Table II. The column A is the visualization of groundtruth poses, the column B in yellow box is our SimpleBaseline+FiDIP model results. The following columns show the results of models ordered based on the in AP accuracy when tested on SyRIP Test500. The columns C to J are: DarkPose:AP=98.0, DarkPose:AP=97.7, SimpleBaseline:AP=97.6, DarkPose:AP=97.4, SimpleBaseline:AP=97.3, DarkPose:AP=95.2, FasterR-CNN:AP=93.4, and FasterR-CNN:AP=91.9. Incorrect predictions are highlighted in red in each image.