

UNSUPERVISED DOMAIN ADAPTATION LEARNING FOR HIERARCHICAL INFANT POSE RECOGNITION WITH SYNTHETIC DATA

Cheng-Yen Yang¹, Zhongyu Jiang¹, Shih-Yu Gu¹, Jenq-Neng Hwang¹, Jang-Hee Yoo²

¹ Department of Electrical and Computer Engineering, University of Washington, USA;

² Electronics and Telecommunications Research Institute, South Korea

ABSTRACT

The Alberta Infant Motor Scale (AIMS) is a well-known assessment scheme that evaluates the gross motor development of infants by recording the number of specific poses achieved. With the aid of the image-based pose recognition model, the AIMS evaluation procedure can be shortened and automated, providing early diagnosis or indicator of potential developmental disorder. Due to limited public infant-related datasets, many works use the SMIL-based method to generate synthetic infant images for training. However, this domain mismatch between real and synthetic training samples often leads to performance degradation during inference. In this paper, we present a CNN-based model which takes any infant image as input and predicts the coarse and fine-level pose labels. The model consists of an image branch and a pose branch, which respectively generates the coarse-level logits facilitated by the unsupervised domain adaptation and the 3D keypoints using the HRNet with SMPLify optimization. Then the outputs of these branches will be sent into the hierarchical pose recognition module to estimate the fine-level pose labels. We also collect and label a new AIMS dataset, which contains 750 real and 4000 synthetic infants images with AIMS pose labels. Our experimental results show that the proposed method can significantly align the distribution of synthetic and real-world datasets, thus achieving accurate performance on fine-grained infant pose recognition.

Index Terms— Infant Pose Recognition, Infant Pose Estimation, Unsupervised Domain Adaptation

1. INTRODUCTION

Autism Spectrum Disorder (ASD) is a developmental disability that can cause significant social, communication, and behavioral challenges. Recent medical research suggests that early signs of ASDs may first manifest within the motor control system and present as a motor delay. According to the studies [1, 2, 3], infants who have delays in acquiring motor skills are at higher risk of developing ASD and may serve as an early indicator of neuro-developmental disorder. To better measure or quantify the level of motor development, Alberta Infants Motor Scale (AIMS) [4] is introduced as a scale

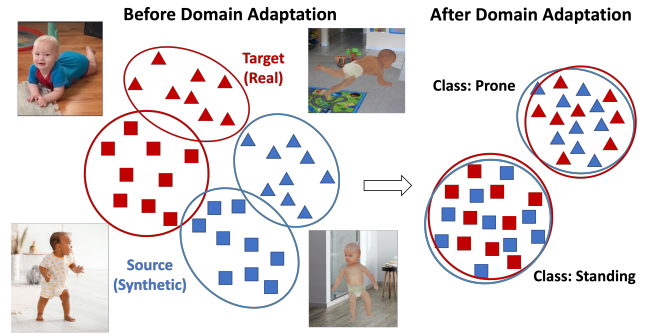


Fig. 1. Unsupervised domain adaptation can exploit the local affinity to capture the fine-grained information and align the distribution accordingly, which can significantly boost the performance of systematic AIMS assessment of real infant images using synthetic infant training data.

assessment procedure to evaluate and track infant motor milestones based on counting the number of gross motor skills that the target can achieve. However, traditional AIMS assessment requires trained professionals to conduct, which are considered time-consuming and inefficient. Moreover, the collection of enough and diversified infant pose images for algorithmic development is challenging due to privacy concerns and institutional Review Board (IRB) regulations. To overcome these issues, a systematic AIMS assessment system, which incorporates the deep learning based 3D infant pose estimation trained by synthetic infant pose data and unsupervised domain adaptation technologies, is proposed to show promising performance for effective AIMS assessment of real-world infant in this paper.

Due to the difficulty in collecting and annotating such fine-grained poses, many datasets collect long untrimmed sequences that lack several distinct poses. Moreover, real-world infant-related data are extremely limited because of privacy concerns and institutional Review Board (IRB) regulations, resulting in most related research on infants using the synthetic datasets for model training. However, one of the problems of training using synthetic data is the domain shift between the synthetic data and real-world data, which

* This work was supported by Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (2019-0-00330).

often leads to performance degradation and poor generalization ability of the trained model. One attempt to solve such a problem is transferring a model learned on a labeled source domain to an unlabeled target domain, known as Unsupervised Domain Adaptation (UDA). Specifically, we have synthetic infant images with pose labels as the source domain and real-world infant images without labels as the target domain. By using UDA, our final goal is to align the source and target domains in the feature space for further applications.

Our contributions are in two-folds: (1) Present a new infant pose dataset with both synthetic and real-world images with fine-level annotation labels, that allow us to evaluate the cross-dataset generalization ability of the model, and (2) Integrate an unsupervised domain adaptation (UDA) algorithm into the hierarchical pose recognition framework to enable transfer learning across domains for better feature extractions of the existing CNN framework.

The paper is organized in the following: We review some related works to our research in Sec. 2. Then we describe our AIMS infant pose dataset synthesizing process in Section 3. In Sections 4 and 5, the methods and experimental results of the infant pose recognition will be presented, followed by the conclusion in Section 6.

2. RELATED WORKS

2.1. Human Pose Estimation & Recognition

Pose Estimation In general, 2D human pose estimation methods can be classified into bottom-up and top-down approaches. Top-down approaches [5, 6, 7] first detect human bounding boxes and then perform human keypoint detection within every bounding box’s region. On the other hand, bottom-up approaches [8, 9, 10] first detect all keypoints on all humans in the image and then assign keypoints belonging to the same person of their owner.

Most recent 3D human pose estimation works take 2D skeletons as input. Bogo et al.[11] adopt the Skinned Multi-Person Linear (SMPL) [12], a statistical body shape model, as an initial human skeleton and optimize the skeleton by minimizing the reprojection error to get the final 3D human pose. Zhao et al.[13] adopt a Graph Convolutional Network (GCN) to process the features of 2D joints as node features to generate 3D joint predictions. Pavllo et al.[14] propose the most well-known temporal-based 3D solutions, the VideoPose3D, which adopt 2D joints of hundreds of frames as input to predict a single 3D skeleton. We take advantage of Skinned Multi-Infant Linear (SMIL)[15], which is an infant version of SMPL, to generate our synthetic dataset, resulting in an image-based infant pose recognition task. SMPLify[11], which minimizes the error between the 3D model joints and detected 2D joints, is adopted for our 3D infant pose estimation.

Pose Recognition Vision-based human pose recognition aims

to obtain posture and predict the corresponding action from input images or video sequences. Most pose or action recognition works are developed closely with pose estimation by leveraging the skeleton-based approaches for prediction [16, 17]. Current publicly available human pose or action datasets are predominantly from scenes such as sports and daily activities performed by adult humans [18, 19] which differ dramatically from the infant motions in terms of achievable poses.

2.2. Unsupervised Domain Adaptation

Various works have been targeting domain adaptation to overcome the domain shift problems. Sener et al. [20] propose to use clustering techniques and pseudo-labels to obtain discriminative features. Taigman et al. [21] propose cross-domain image translation methods. Ganin et al. [22] propose a representative method of distribution matching involving training a domain classifier using the intermediate features and generating the features that deceive the domain classifier. This method utilizes the similar techniques used in generative adversarial networks (GANs). The category classifier is trained to predict the task-specific category labels. And the domain classifier is trained to predict the domain of each input. The two classifiers share feature extraction layers that are trained to predict the label of source samples correctly and deceive the domain classifier. Thus, the distributions of the intermediate features of the target and source samples are made similar. However, an issue of unsupervised domain adaptation by back-propagation is that the target features can be near a task-specific classifier’s boundary, which will cause the target samples far from source ones (ambiguous features) to be likely misclassified after alignment.

2.3. Infant Dataset

Privacy remains one of the issues for infant-related dataset collection process. Therefore, the majority of the existing infant datasets are synthetic images. Currently, there are only limited infant-related datasets: MINI-RGBD [23], SyRIP [24], and Zhou et al. [25]. MINI-RGBD mapped real infant movements to the SMIL model, generating RGB and depth video sequences with 2D and 3D joint coordinates. However, these data are synthesized from infants under seven months old and thus present simple poses with small changes over samples. SyRIP is composed of two portions, real and synthetic. The real part consists of 700 infant images collected from public sources like Youtube and Google, while the synthetic part consists of 1000 images rendered using the SMIL model. The 2D joint coordinates are fully annotated in COCO format [26] for these images, which is a huge contribution toward the infant pose estimation research but lack labels for the pose classification task. Zhou’s dataset contains 5500 synthetic images with 11 classes selected from AIMS, but the in-

Table 1. The selected infant poses in AIMS [4] in our work with 4 coarse-level and 12 fine-level labels.

Coarse-Level	Fine-Level
Prone	Prone Lying
	Forearm Support
	Reciprocal Crawling
	Four-Point Kneeing
Supine	Supine Lying
	Hands to Knee/Feet
	Rolling
Sitting	Sitting w/ Support
	Sitting w/ Arm Support
	Sitting w/o Support
Standing	Four-Point Standing
	Standing

sufficient corresponding real-world evaluation portion makes it difficult to justify the model’s generalization ability.

3. AIMS DATASET

To provide an evaluation of the generalization ability of the models, we need a dataset that includes pose labels of the real-world infant image. We start from a small number of real-world infant samples and use the SMIL-based model to enlarge our real-world infant dataset with labeled synthetic data.

3.1. Skinned Multi-Infant Linear (SMIL) Model

Skinned Multi-Person Linear Body (SMPL) [12] is a skinned vertex-based model that is able to represent different human body shapes and poses with the parameters learned from data. Skinned Multi-Infant Linear body model (SMIL) [15] is a derived version of the SMPL learned from the sequences of freely moving infants in [23]. Some pose coefficients $\theta \in \mathbb{R}^{3 \times N_j}$ and the shape coefficients $\beta \in \mathbb{R}^{N_s}$ serve as input, and the output is a mesh consisting of $N_v = 6890$ vertices with $N_j = 24$ joints defined in SMPL.

In order to obtain realistic pose and shape parameters for infants, we take advantage of SMPLify [27] to better fit any given infant image by minimizing the overall loss function:

$$L(\theta, \beta) = L_{J_{2D}} + L_\theta + L_\beta, \quad (1)$$

where $L_{J_{2D}}$ denotes the distance between estimated 2D joints and the 2D projection of the 3D joints. L_θ and L_β respectively denotes the simple L_2 prior for body pose and body shape.

3.2. Rendering

After fitting SMIL model for each infant instance in the image, we can then render synthetic images using different tex-

ture of the infant model, different backgrounds, and reasonable translation operations. The ground-truth 3D keypoint coordinates \mathbf{X}_{3d} can be obtained directly from the fitted SMIL model, where all 3D keypoints will be normalized with respect to the distance from nose to pelvis to ensure the scaling consistency.

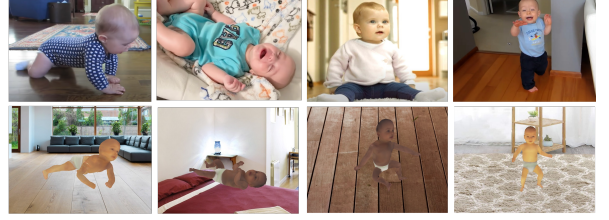


Fig. 2. Sample images from our collected AIMS dataset. The first row is from the real portion and the second row is the synthetic portion.

As for 2D keypoint coordinates, we need to reproject the 3D keypoints in the world coordinates to the image coordinates based on the camera parameters used to render each synthetic sample. The ground-truth 2D keypoint coordinates \mathbf{x}_{2d} can thus be obtained by:

$$\mathbf{x}_{2d} = \mathbf{K} \cdot [\mathbf{R}|\mathbf{T}] \cdot \mathbf{X}_{3d}, \quad (2)$$

where \mathbf{K} and $[\mathbf{R}|\mathbf{T}]$ are the pre-defined camera intrinsic and extrinsic parameters.

We extend the real portion of the SyRIP dataset [24] by annotating and categorizing the real infant portion into 12 selected fine-level gross motor poses, a very small portion ($\approx 5\%$) of samples are withdrawn due to the poses not falling into any of defined fine-level poses. We randomly assign different camera parameters and remove those unnatural samples after syntheses. In addition, we also collect 200 background images from Google, and select 1000 scenes from INDOOR dataset [28] under the labels like bedroom, children room, and nursery, to mimic the real-world data and prevent overfitting on the synthetic images.

In total, the entire synthetic portion consists of around 4000 samples, while the real portion consists of 750 samples with ground-truth 2D keypoints (image coordinate), 3D keypoints (world coordinate), coarse-level and fine-level AIMS labels.

4. METHODOLOGY

4.1. Image Branch: Domain Adaptation Network

For image-level classification, the gap between real-world infant samples and synthetic infant samples often causes performance degradation and leads to inaccurate predictions. We can formulate the infant pose recognition task as an unsupervised domain adaptation scenario. More specifically, given a

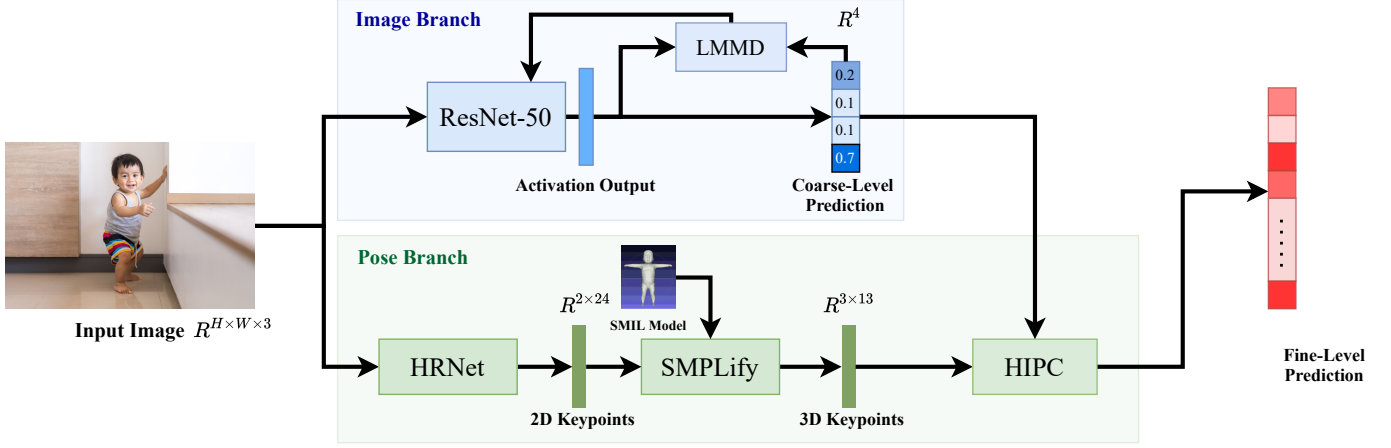


Fig. 3. The overview of pipeline of our proposed infant action recognition with unsupervised domain adaptation learning.

source domain $D_s = \{(\mathbf{x}^s, \mathbf{y}^s)\}$ and an unlabeled target domain $D_t = \{\mathbf{x}^t\}$, which are sampled from different distributions s and t , respectively. The goal of unsupervised domain adaption (UDA) is to train the CNN classifier that reduces the discrepancy of the two distributions. Specifically, we adopt Local Maximum Mean Discrepancy (LMMD) to aid our UDA on aligning the relevant subdomain distributions of domain specific layer activation across different domains.

The unbiased estimator of LMMD can be expressed as:

$$d_{\mathcal{H}}(s, t) = \frac{1}{n_c} \sum_{c \in C} \left\| \sum_{x^s \in D_s} w_c^s \phi(x^s) - \sum_{x^t \in D_t} w_c^t \phi(x^t) \right\|_{\mathcal{H}}^2, \quad (3)$$

where \mathcal{H} is the reproducing kernel Hilbert space (RKHS) [29] with kernel k . The kernel l represents $k(x^s, x^t) = \langle \phi(x^s), \phi(x^t) \rangle$, which is the inner product of two vectors of some feature mapping operation $\phi(\cdot)$.

The important characteristic of LMMD is the w_c^s and w_c^t , which denote the weights of x belonging to a given class c in the source and target domains, respectively. The weights for source domain w_c^s can be easily computed using the ground truth label as an one-hot vector $w_c^s = y_i^{sc} / \sum_{i \neq j} y_j^{sc}$. However for the unlabeled target domain, the weights w_c^t have to be adaptively estimated using the output z of each activation layer $l \in L$ in order to compute Eq. 3.

Finally, the adaptation loss will be multiplied by a coefficient λ and added to the classification loss, which is the naive cross-entropy loss. The overall loss to be minimized becomes:

$$\mathcal{L}_{overall} = \mathcal{L}_{classify} + \lambda \cdot \mathcal{L}_{adapt}. \quad (4)$$

4.2. Pose Branch: 2D/3D Pose Estimation

We adopt HRNet[7] and SMPLify[11] for our 2D and 3D pose estimation. For 2D Pose Estimation, HRNet takes a single

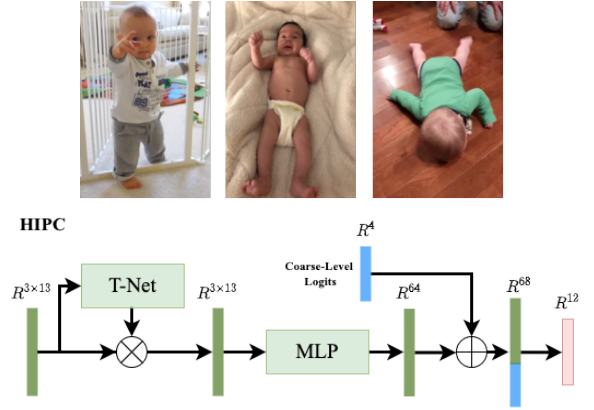


Fig. 4. An example of different poses that result in the same 2D keypoints in different view angle and the detailed architecture of HIPC aiming to alleviate such problem by using the logits from the image branch.

image I as input, and generates heatmaps H for all the keypoints. The coordinates \mathbf{x}_{2d} can be obtained by finding points with the highest values in H . After getting \mathbf{x}_{2d} , SMPLify[11] is used to optimize the SMIL model iteratively by minimizing the reprojection error of projected \mathbf{X}_{3d} with \mathbf{x}_{2d} .

4.3. HIPC: Hierarchical Infant Pose Classifier

For infant pose recognition, to overcome the confusion caused by the viewing perspectives, 3D skeletons should be used for better recognition performance. However, as Fig 5, unlike adult human pose recognition, the view angles of infant images are more flexible, and two similar 3D skeletons may lead to two completely different fine level poses. As a result, we take advantage of a Hierarchical Infant Pose Classifier (HIPC)[25], which takes the 3D keypoints \mathbf{X}_{3d} and coarse-level recognition logits as input, for getting better fine level recognition result.

Table 2. Experimental results of our proposed method for coarse-level and fine-level classification. The **bold** text denotes the highest top-1 accuracy achieved under the same experiment setting.

Task	Method	Top-1 Accuray
Coarse-Level Classification	Image Branch	75.5%
	Image Branch (w/ LMMD)	85.3%
	Pose Branch	83.3%
Fine-Level Classification	Image Branch	40.4%
	Image Branch (w/ LMMD)	50.5%
	Pose Branch	68.0%
	Ours	76.8%

Table 3. Effect of domain adaptation loss when using different training dataset (i.e., different source domain distribution).

Source Domain	Method	Top-1 Accuracy (Coarse-Level)
<i>Synthetic</i>	Image Branch	45.3%
	Image Branch (w/ LMMD)	60.3%
<i>Synthetic+Real</i>	Image Branch	75.5%
	Image Branch (w/ LMMD)	85.3%

5. EXPERIMENTAL RESULTS

To evaluate the performance of both coarse-level and fine-level infant pose recognition tasks, we train the model using all of the synthetic dataset along with a subset of the ground truth labelled real dataset, where 4 coarse labels and 12 fine sub-labels with a total of 4000 samples are used. We evaluate our model using the test subset of the real data with a total of 198 images and record the Top-1 accuracies. The codes are implemented in Pytorch and we conduct the experiments on one Nvidia GeForce GTX Titan XP card.

5.1. Performance of Infant Pose Recognition

The experimental results are shown in Table 2, which shows the top-1 accuracies for coarse-level pose classification (4 classes) and fine-level pose classification (12 classes) based on the same algorithmic configurations. With the help of adding LMMD loss and a λ of 0.5, the domain adaptation module introduced to the naive ResNet-50 is able to improve the performance of our image-branch coarse-level classification from 75.5% to 85.3% (using *Syn+Real*). Moreover, the fine-level classification results can be improved from 68.0% to 76.8% after using the logits from the image-branch to guide the prediction from pose branch. In Fig. 5, we can clearly see that most of those misclassified samples across coarse-levels (i.e., outside of the gray bounding boxes) had been corrected with the aid of logits from image-branch in our proposed method.

Our best model leverages both unsupervised domain adaptation for coarse-level classification and hierarchical pose

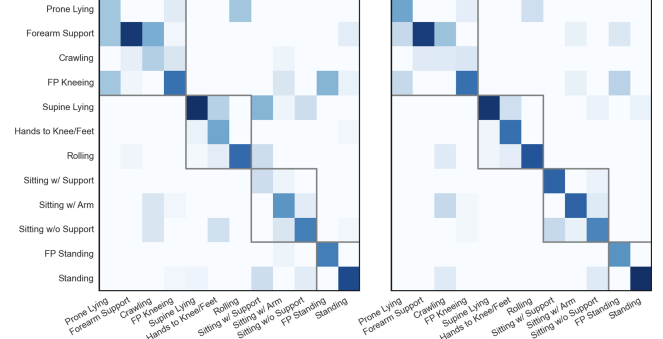


Fig. 5. A comparison of the confusion matrices for fine-level classification results from baseline (left) and our purposed method (right). The gray boxes represent the fine-level labels with the same coarse-level label.

recognition framework and can truly overcome the source-target domain distribution mismatch, as we achieved a 76.8% accuracy on fine-level classification.

5.2. Ablation Study

We now analyze the effect of domain adaptation on different training dataset. The source domain denotes the dataset used during training is denoted as *Synthetic*, which represents the entire synthetic portion from AIMS dataset and *Synthetic+Real* represents the setting of adding a small number of real-world infant samples into the source domain. As shown in Table 5, LMMD loss can significantly improve the performance of classification without any major modification being made to the CNN-based backbone, since the adaptation loss is computed using the activation outputs and pseudo-labels only (i.e., from 45.3% to 60.3% Top-1 accuracies on *Synthetic* and from 75.5% to 85.3% Top-1 accuracies on *Synthetic+Real*).

6. CONCLUSION

Fine-level infant pose recognition may help doctors or parents determine infants' motor skill development. In this paper, we proposed a new AIMS Dataset for fine-level infant pose recognition, with both synthetic and real-world data. With this dataset, we integrated an unsupervised domain adaptation algorithm into the hierarchical pose recognition framework to enable transfer learning across domains for better feature extractions of existing CNN framework, and finally, we achieved 76.8% Top-1 accuracy with the domain-adopted model on the AIMS test dataset.

7. REFERENCES

- [1] Anjana N. Bhat, Rebecca J. Landa, and James C. (Cole) Galloway, "Current Perspectives on Motor Functioning in Infants, Children, and Adults With Autism Spectrum Disorders," *Physical Therapy*, vol. 91, no. 7, pp. 1116–1129, 07 2011.

- [2] Joanne Flanagan, Rebecca Landa, Anjana Bhat, and Margaret Bauman, "Head lag in infants at risk for autism: A preliminary study," *The American journal of occupational therapy*, vol. 66, pp. 577–85, 09 2012.
- [3] James Patterson, Vickie Armstrong, Eric Duku, Annie Richard, Martina Franchini, Jessica Brian, Lonnie Zwaigenbaum, Susan Bryson, Lori-Ann Sacrey, Caroline Roncadin, and Isabel Smith, "Early trajectories of motor skills in infant siblings of children with autism spectrum disorder," *Autism Research*, 11 2021.
- [4] Piper Martha and Darrah Johanna, *Motor Assessment of the Developing Infant*, Saunders, 1994.
- [5] Shih-En Wei, Varun Ramakrishna, Takeo Kanade, and Yaser Sheikh, "Convolutional pose machines," in *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 2016, pp. 4724–4732.
- [6] Alejandro Newell, Kaiyu Yang, and Jia Deng, "Stacked hourglass networks for human pose estimation," in *European conference on computer vision*. Springer, 2016, pp. 483–499.
- [7] Jingdong Wang, Ke Sun, Tianheng Cheng, Borui Jiang, Chaorui Deng, Yang Zhao, Dong Liu, Yadong Mu, Mingkui Tan, Xinggang Wang, et al., "Deep high-resolution representation learning for visual recognition," *IEEE transactions on pattern analysis and machine intelligence*, 2020.
- [8] Zhe Cao, Gines Hidalgo, Tomas Simon, Shih-En Wei, and Yaser Sheikh, "Openpose: realtime multi-person 2d pose estimation using part affinity fields," *IEEE TPAMI*, vol. 43, no. 1, pp. 172–186, 2019.
- [9] George Papandreou, Tyler Zhu, Liang-Chieh Chen, Spyros Gidaris, Jonathan Tompson, and Kevin Murphy, "Personlab: Person pose estimation and instance segmentation with a bottom-up, part-based, geometric embedding model," in *ECCV*, 2018, pp. 269–286.
- [10] Bowen Cheng, Bin Xiao, Jingdong Wang, Honghui Shi, Thomas S. Huang, and Lei Zhang, "Bottom-up higher-resolution networks for multi-person pose estimation," *CoRR*, vol. abs/1908.10357, 2019.
- [11] 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 *ECCV*, 2016, pp. 561–578.
- [12] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black, "SMPL: A skinned multi-person linear model," *ACM Trans. Graphics (Proc. SIGGRAPH Asia)*, vol. 34, no. 6, pp. 248:1–248:16, Oct. 2015.
- [13] Long Zhao, Xi Peng, Yu Tian, Mubbassir Kapadia, and Dimitris N. Metaxas, "Semantic graph convolutional networks for 3d human pose regression," in *Proceedings of the IEEE/CVF CVPR*, June 2019.
- [14] 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 CVPR*, June 2019.
- [15] Nikolas Hesse, Sergi Pujades, Javier Romero, Michael J. Black, Christoph Bodensteiner, Michael Arens, Ulrich G. Hofmann, Uta Tacke, Mijna Hadders-Algra, Raphael Weinberger, Wolfgang Müller-Felber, and A. Sebastian Schroeder, "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 (MICCAI)*. Springer, 2018.
- [16] Limin Wang, Yuanjun Xiong, Zhe Wang, Yu Qiao, Dahua Lin, Xiaoou Tang, and Luc Van Gool, "Temporal segment networks: Towards good practices for deep action recognition," in *ECCV*, 2016.
- [17] Haodong Duan, Yue Zhao, Kai Chen, Dian Shao, Dahua Lin, and Bo Dai, "Revisiting skeleton-based action recognition," *CoRR*, vol. abs/2104.13586, 2021.
- [18] Mykhaylo Andriluka, Leonid Pishchulin, Peter Gehler, and Bernt Schiele, "2d human pose estimation: New benchmark and state of the art analysis," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2014.
- [19] Sam Johnson and Mark 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.
- [20] Ozan Sener, Hyun Oh Song, Ashutosh Saxena, and Silvio Savarese, "Learning transferrable representations for unsupervised domain adaptation," in *NIPS*, 2016.
- [21] Yaniv Taigman, Adam Polyak, and Lior Wolf, "Unsupervised cross-domain image generation," *ArXiv*, vol. abs/1611.02200, 2017.
- [22] Muhammad Ghifary, W. Bastiaan Kleijn, Mengjie Zhang, David Balduzzi, and Wen Li, "Deep reconstruction-classification networks for unsupervised domain adaptation," *CoRR*, vol. abs/1607.03516, 2016.
- [23] 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 *ECCV 2018 Workshops*. 2018, Springer International Publishing.
- [24] Xiaofei Huang, Nihang Fu, Shuangjun Liu, and Sarah Ostadabbas, "Invariant representation learning for infant pose estimation with small data," in *IEEE International Conference on Automatic Face and Gesture Recognition (FG)*, 2021, December 2021.
- [25] 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)*, 2021, pp. 1340–1344.
- [26] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, Lubomir D. Bourdev, Ross B. Girshick, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick, "Microsoft COCO: common objects in context," *CoRR*, vol. abs/1405.0312, 2014.
- [27] 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 *ECCV 2016*. Oct. 2016, Lecture Notes in Computer Science, Springer International Publishing.
- [28] Ariadna Quattoni and Antonio Torralba, "Recognizing indoor scenes," in *2009 IEEE Conference on Computer Vision and Pattern Recognition*, 2009, pp. 413–420.
- [29] S.K. Zhou and R. Chellappa, "From sample similarity to ensemble similarity: probabilistic distance measures in reproducing kernel hilbert space," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 28, no. 6, pp. 917–929, 2006.