MONOCULAR 3D HUMAN POSE ESTIMATION WITH DOMAIN FEATURE ALIGNMENT AND SELF TRAINING

Yan-Hong Zhang, Calvin Ku, Min-Chun Hu and Hung-Kuo Chu

yanhongzhang212@gmail.com; calvinku1209@gmail.com;
anitahu@cs.nthu.edu.tw and hkchu@cs.nthu.edu.tw
National Tsing Hua University, Taiwan

ABSTRACT

Despite great success in 3D monocular human pose estimation, the progress of accurate prediction for unseen poses or complex backgrounds is still limited due to the lack of labeled data. In this paper, we use synthetically generated images with 3D ground truth and unlabelled real data to address this domain gap challenge. Unlike recent works that apply the adversarial loss to their models, we propose a novel domain feature alignment method (DFA) that avoids the disadvantages of unstable training and wrong alignment. In addition, our method leverages self-training with data enhancement to create robust pseudo-labels for real data. The experimental results show the effectiveness of combining self-training with our DFA method on Human 3.6M testing data without using any 3D ground truth real data.

1. INTRODUCTION

Human pose estimation based on a single RGB image has been widely used in many applications such as sports performance analysis, human-computer interaction, human action recognition, and human tracking in autonomous cars.Driven by powerful deep learning technologies, 2D human pose estimation has made a great progress in recent years, but accurate 3D human pose prediction is still challenging due to the lack of a labeled dataset that contains a wide range of poses, appearances, occlusions, and backgrounds. Many existing methods have good performance only on specific public datasets, but it remains a challenge to apply existing data-oriented deep learning models to commercial uses directly, especially when the application requires accurate 3D body joint prediction . Currently, 3D data collection is costly, time-consuming, and labor-intensive.

In order to solve the complexity and cost of 3D data collection, recent research has begun to use computer graphics technology to generate synthetic data for training deep learning models. Using synthetic data allows researchers to generate a large amount of annotated data with a high degree of controllability, like creating various scenes, characters' actions, and characters' appearance. However, there is a domain gap of the image texture between the synthetically generated data and the real world image, and models trained with synthetic data do not generalize well in real data.

Currently, there are two main approaches for solving the domain adaptation problem: adversarial learning and self-training based. Adversarial learning uses discriminators to align the domain feature by optimizing the feature extractor to produce domain invariant features. However, the adversarial-based method suffers from unstable training due to the pair training requirement. Therefore such methods might fail if the postures of the paired inputs are drastically different. On the other hand, the self-training based method can use a confidence threshold to filter possible wrong predictions and generate more reliable pseudo labels. Nevertheless such method lacks a mechanism for domain feature alignment and is not effective if the pseudo label generator is not strong enough.

Therefore, in this work, we propose a domain feature alignment (DFA) method with a self-training mechanism, which utilizes the advantages of self-training and avoids the disadvantages of using discriminators. The decoder of our heatmap prediction model is only trained in the source domain to reconstruct the correct 3D pose, and therefore it forces the encoder to produce domain invariant features, which acts like a discriminator in the adversarial learning method. In addition, to solve the problem of pseudo label generation, we integrate a self-training mechanism with enhanced pseudo-labeled data generation [1] that can generate more reliable pseudo labels. To the best of our knowledge, we are among the first to combine domain adaptation with self training architecture for 3D human pose estimation.

2. RELATED WORK

2.1. Unsupervised Domain Adaptation

Unsupervised domain adaptation has been well studied for many object detection and semantic segmentation tasks. Several methods try to minimize the appearance domain through the image to image translation method [2, 3, 4]. Another line of work studies how to cross the domain gap by designing and minimizing the feature measurement metric, such as

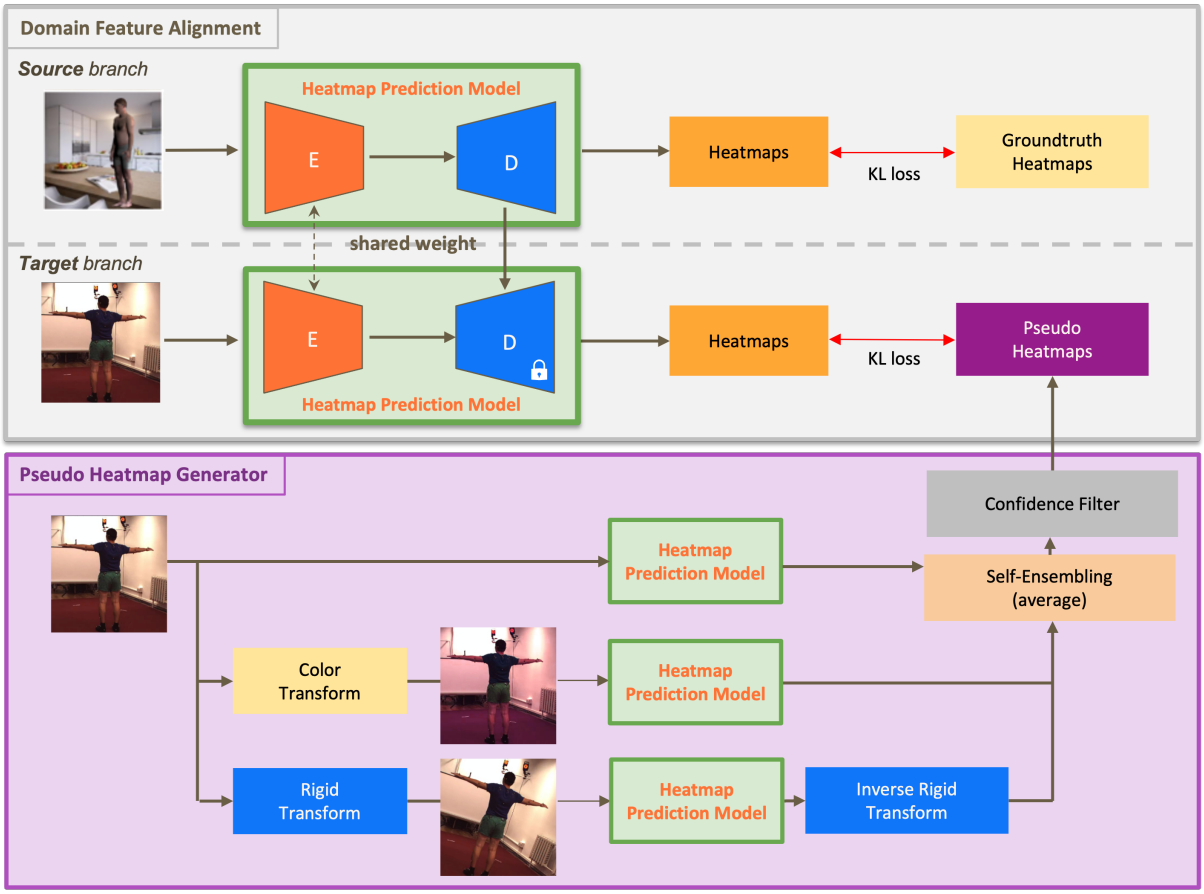

Fig. 1: Proposed learning architecture with our feature alignment mechanism, where the decoder of the target branch is locked.

maximum mean discrepancy [5] or CORAL [6]. Some methods apply adversarial loss based on a domain discriminator to align the source domain feature with the target domain feature [7, 8, 9]. [10, 11] try to solve the domain adaptation problem according to theories [12]. Currently, there are only a few methods that deal with the 3D human pose problem using domain adaptation methods. Existing works try to apply multiple joint domain discriminator [13] or leverage the depth and body part segmentation information to minimize the domain gap [14]. Different from prior works, our method explores a new mechanism to implicitly align the source domain and the target domain by freezing the decoder in the adaptation stage Section 3.2 and achieves better PA-MPJPE than current state-of-the-art methods.

2.2. Self-training

Self-training methods are commonly used in semi-supervised deep learning tasks [15], such as object detection and segmentation. Most self-training models would generate new pseudo-labels after every training step until meeting the end conditions. The biggest challenge in this field is that the generated pseudo-labels are usually noisy and thus lead to unstable training process. A number of regularizing works were proposed [16, 17] to solve this issue. While other works design mechanisms to filter out noisy pseudo-labels [7]. Mu et al. introduce a self-ensemble method for 2D animal skeleton prediction [1]. In addition to learning a prediction model based on the original training data, they produce new training images in different ways (e.g. applying color transform or rigid transform on the original one) and compute the average output of these result for the final pseudo-label, which enables the model to neglect noisy inputs. They also introduce an automatic confidence selection mechanism, which allows the model to initially learn from high confidence samples and gradually learn more complex data by choosing samples with lower confidence. In this work, we incorporate their idea into our domain adaptation alignment framework to avoid the feature misalignment between the source and target domains.

3. APPROACH

In this paper, we aim to solve the unsupervised domain adaptation (UDA) problem for 3D human pose estimation. We follow the notation from [12] and define a domain as an ordered pair consistency of a distribution D on input space X to

the label function $f : X \longrightarrow Y^K$ that maps X to the label space Y^K, where K is the number of key points for each input. The source domain and target domain are denoted by $\langle D_s, f_s \rangle$ and $\langle D_t, f_t \rangle$, respectively. In UDA, the model is trained on labeled data from the source domain and the unlabeled data from the target domain. The goal of our problem is to find a function that minimizes the error on both D_s and D_t without using any manually labeled data from the target domain.

As illustrated in Figure 1, we propose a self-training 3D human pose estimation system, which comprises three main stages, i.e., *pretrain stage*, *pseudo heatmap generation stage*, and *adaptation stage*. In the *pretrain stage*, we train the source-domain branch of the heatmap prediction model with the labeled synthetic data. In the *pseudo heatmap generation stage* (Section 3.1), the model trained in the previous stage is used to generate *pseudo heatmaps* for unlabeled real data. The system repeats the *pseudo heatmap generation stage* and the *adaptation stage* to obtain the final heatmap prediction.

Algorithm 1: Pseudo-Label Generation Algorithm

Input : Target dataset X_t; model trained in the previous time-step $f^{(k-1)}$

Intermediate Result: P_α, P_β are predictions after applying invariance and equivalence transform.

Output : Pseudo-label $\hat{Y}_t^{(t)}$; confidence score $C_{tar}^{(k)}$

for X_{tar}^i *in* X_{tar} **do**
  $P_\alpha = f^{(k-1)}(T_\alpha(X_{tar}^i))$ `// Invariance transform`
  $P_\beta = T_\beta^{-1}(f^{(k-1)}(T_\beta(X_t^i)))$ `// Equivalence transform`
  Ensemble P_α and P_β to get $(\hat{Y}_t^n, C_t^{(k)})$
end
Sort $C_t^{(k)}$ and obtain C_{thresh} based on a fixed policy.
Set $C_t^{(k),i} = \mathbb{1}(C_t^{(k),i} \geq C_{thresh}), \forall i$

3.1. Pseudo Label Generation

Conventionally, self-training methods directly use the model trained in the previous stage to generate the pseudo labels for the unlabeled data. To create more reliable pseudo labels that can be used for the adaptation stage, a self-ensembling mechanism is applied in our pseudo heatmap generator. To be more precise, color transformation and rigid transformation are applied on the target images, and the model trained in the previous stage is applied to predict the heatmaps given the original real image and the transformed images as the inputs. By ensembling the prediction outputs of the original image and the transformed images, the generated pseudo labels would be more robust to the change of environmental lighting and viewing angle.In this work, we simply averaging the three prediction outputs. Moreover, unreliable pseudo labels are filtered out so that the model can learn with easier samples first in the *adaptation stage*. With the increase of training iteration, the threshold of confidence filtering is loosened and the model can learn with more complex samples. The proposed pseudo label generation stage is described in Algorithm 1.

3.2. Domain Feature Alignment (DFA)

Our self-training method lets the model generate a better sample of new pseudo heatmaps for the next training stage and improves the training performance. The *pretrain stage* is trained on the source domain with synthetic data, which guides the pseudo heatmaps generation module to generate more reliable heatmaps for the *adaptation stage*. In the *adaptation stage*, the model is trained based on both the synthetic data with ground truth heatmaps and the real data with the pseudo heatmaps.

Traditional domain adversarial training like [7] requires a discriminator that forces the encoder to align the source and target features. However, this kind of methods do not perform well if the source and target poses are vert different. Therefore, instead of using a discriminator, we propose a domain feature alignment method which freezes the decoder of the target branch. In this way, the decoder will force the encoder to align the source and target features so that the decoder can reconstruct the correct 3D pose. The loss function $L^{(k)}$ for the k^{th} iteration is defined by the Kullback–Leibler divergence loss of both the source heatmaps (s) and the target heatmaps (t), as in Equation 1.

$$L^k = \alpha \sum_i L_{KL}(f^k(X_s^i), Y_s^i) + \beta \sum_j L_{KL}(f^k(X_t^j), \hat{Y}_t^{(k-1),j}) \quad (1)$$

4. EXPERIMENTS

4.1. Datasets

For all the experiments, we use SURREAL [18] as our source domain dataset while using images from Human 3.6M [19] without any ground truth as our target domain dataset. Moreover, MPII [20] dataset is applied as an additional source domain dataset for the *pretrain stage*. We will compare the results of using only SURREAL dataset and using both SURREAL and MPII datasets as the source domain data.

SURREAL is a large-scale dataset containing 6 million frames of synthetic humans in photo-realistic images under large variations in shape, texture, view-point and poses. The synthetic humans are rendered based on 3D sequence of human motion capture data and the SMPL body model, whose parameters are fit by the MoSh method given raw 3D MoCap

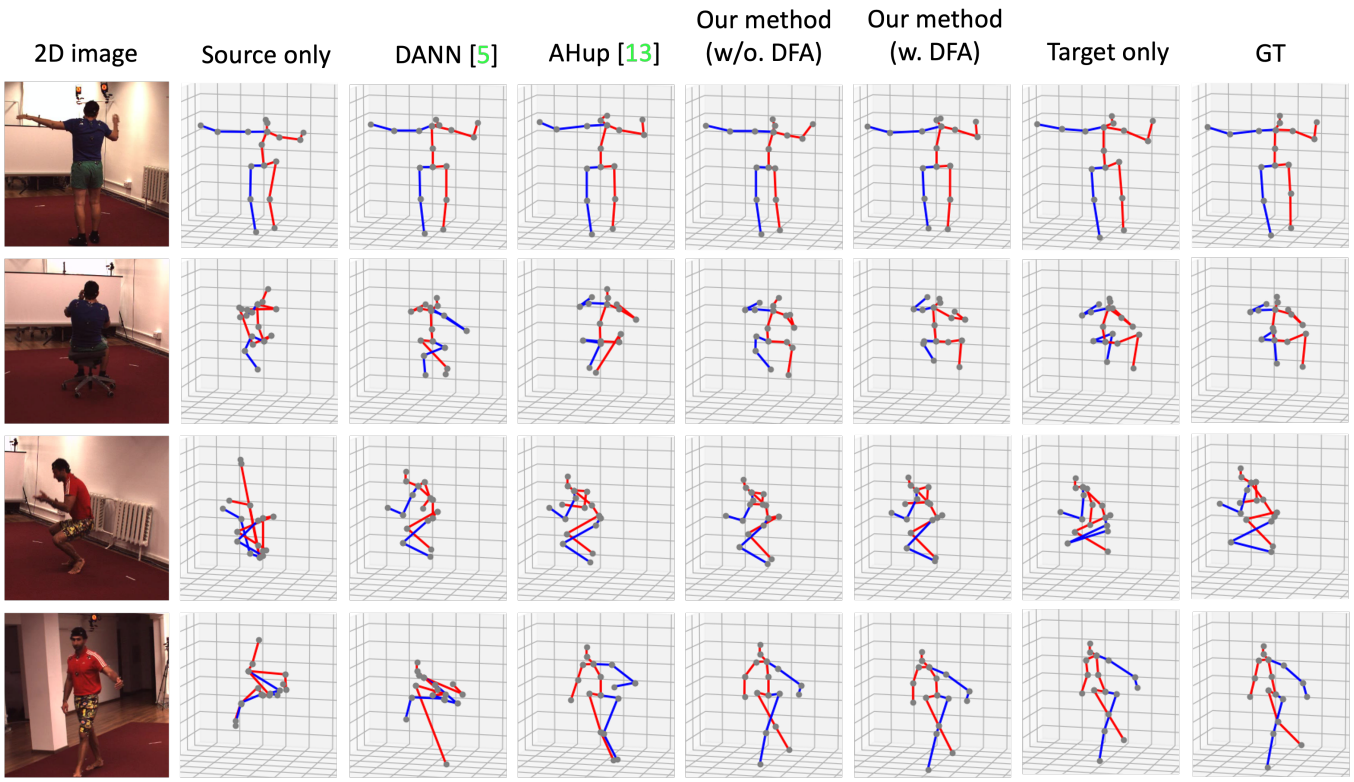

Fig. 2: 3D visualization comparison of our method with other state of the arts without pretraining on MPII.

marker data. **Human 3.6M dataset (H3.6M)** is a commonly used dataset for training and evaluating 3D human pose estimation. It consists of 3.6 million 3D poses that is captured from 4 different view points of 5 female and 6 male subjects performing common actives like walking, greeting, talking on the phone and eating. **MPII Human Pose** dataset is a state of the art benchmark for evaluation of 2D human pose estimation with around 25000 labeled 2D poses of over 40000 subjects. Each image is extracted from a Youtube video that covers 410 human activities. Since there are no corresponding ground truth of z coordinate, we only consider x and y coordinates to calculate the loss function.

4.2. Implementation Details

The backbone network architecture for our model is based on [21] with input image size of 256x256 pixels. The heatmap prediction model outputs three marginal heatmaps (in the resolution of 32x32 pixels) for the 3D human pose. We use Stochastic Gradient Descent optimizer for the training of both the *pretrain stage* and the *adaptation stage*. In the *pretrain stage*, we set the batch size to 32, learning rate as 0.1, α as 0.5, and epoch number as 150. In the *adaptation stage*, we set the batch size to 8, learning rate as 0.1, α as 0.05, and β as 0.1. The *adaptation stage* is repeated for 20 rounds with 2 epochs for each round.

4.3. Evaluation Result

We use two standard protocols to train and evaluate our results on H3.6M. In protocol 1, the training set consists of subject 1, 5, 6, 7 and 8, and subject 9 and 11 are hold out for validation [19]. The reported error is in mean per joint position error (MPJPE), which is the average euclidean distance of all the 17 estimated 3D joints to the ground truth joints. The same training and validation set of data was used by another error metric, Procrustes Aligned MPJPE (PA-MPJPE) [22]. It is evaluated by calculating MPJPE after rigidly aligning the predicted pose with the ground-truth.

Table 1 shows the quantitative results based on MPJPE and PA-MPJPE. We compared our method with other state of the art methods for 3D human domain adaptation. Two pretrain settings (with and without additional MPII data as the source data) are compared in Table 1. Our proposed model outperforms other works on H3.6M testing data when we do not use additional real data from MPII for the *pretrain stage*. To compare with [14], MPII images were added as part of the source domain data during pretraining. The depth prediction for MPII pose was neglected. In this setting, we were able to achieve the best result in PA-MPJPE, while UDA-Pose [14] has better performance than our method in MPJPE. However, since the authors of UDA-Pose do not provide the details of how to convert UBC3V skeleton joints to H3.6M

Table 1: Detailed results on Human3.6M following Protocol 1 and 2. The error is in millimeters(mm). Top: Without pretraining MPII data. Bottom: Pretrained with MPII data.

Method	Training Dataset	Testing Dataset	MPJPE	PA-MPJPE
w/o MPII				
DANN [5]	Surreal + H36M	H36M	250.0	153.5
Adapted human pose [13]	Surreal + H36M	H36M	132.1	87.2
Ours	Surreal + H36M	H36M	**100.2**	**72.0**
Ours (w/o domain adaptation alignment)	Surreal + H36M	H36M	127.1	87.4
Ours (Only Target)	H36M	H36M	81.2	61.6
Ours (Only Source)	Surreal	H36M	407.4	242.4
w/ MPII				
DANN [5]	Surreal + H36M + MPII	H36M	113.0	79.2
AHup [13]	Surreal + H36M + MPII	H36M	107.0	66.2
UDA-Pose [14]	Surreal + H36M + MPII	H36M	**78.5**	68.9
Ours	Surreal + H36M + MPII	H36M	93.6	**64.7**
Ours (w/o DFA)	Surreal + H36M	H36M	106.7	74.6
Ours (Only Target)	H36M + MPII	H36M	62.7	49,7
Ours (Only Source)	Surreal + MPII	H36M	115.6	82.0

Table 2: Effect of DFA with different settings

Policy	*MPJPE*	*PA_MPJPE*
w/ MPII		
(60% , 0%, 60%)	110.1	79.2
(80% , 0%, 80%)	106.6	76.4
(100%, 0%, 100%)	102.1	73.2
(60% , 5%, 99%)	**100.2**	**72.0**
w/o MPII		
(60% , 0%, 60%)	95.8	66.4
(80% , 0%, 80%)	95.3	66.0
(100%, 0%, 100%)	**93.4**	**64.6**
(80% , 5%, 99%)	93.6	64.7

skeleton joints, we directly use the evaluation results reported in their paper for comparison. The inconsistent definition of joint conversion between their method and ours might result in unfair comparison. In addition, we also examine the upper and lower bound of our work by training with the labeled target (real data) domain and by training with only the labeled source (synthetic data) domain, respectively. Some comparison of the predicted results are visualized in Figure 2.

4.4. Ablation Studies

We conducted ablation studies on the proposed domain feature alignment (DFA) method, which freezes the decoder for the target domain during the training of the *adaptation stage*. As shown in Table 1, we observe that our model with domain feature alignment improves significantly by 26.9mm in MPJPE and 14.6mm in PA-MPJPE when the pretraining setting is without using MPII. When the pretraining setting is with MPII, DFA can also improve the prediction performance by 13.1mm in MPJPE and 9.9mm in PA-MPJPE. We also visualize the effect of domain feature alignment in Figure 2.

We also conducted ablation study on the confidence filter of the Pseudo heatmap generator. As shown in Table 2. The confidence filter involves three percentage variables: *initial*, *step* and *maximum*. After ranking the confidence of the Pseudo heatmaps, the filter chooses the top *initial* percentage data for the training of the next *adaptation stage* and increases the percentage after each round by *step* amount until it reaches the *maximum*. In doing so, our model can learn easier samples first and gradually learn harder samples.

5. CONCLUSION

We have introduced a domain feature alignment method that does not use an additional discriminator to achieve domain alignment. Our domain feature alignment method forces the encoder to take into account the features of the source domain when extracting the features of the target domain, which aligns the target domain features with the source domain features. In addition, we design a self-training method, which involves a pseudo label generator that can reduce the chance of learning the wrong samples and improve predictions for every training round.

6. ACKNOWLEDGEMENT

This research was supported by the Ministry of Science and Technology (Contract MOST 108-2221-E-007-106-MY3, 109-2221-E-007-095-MY3, 110-2221-E-007-061-MY3, and 110-2221-E-007-060-MY3), Taiwan.

7. REFERENCES

[1] Jiteng Mu, Weichao Qiu, Gregory D Hager, and Alan L Yuille, "Learning from synthetic animals," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 12386–12395. 1, 2

[2] Judy Hoffman, Eric Tzeng, Taesung Park, Jun-Yan Zhu, Phillip Isola, Kate Saenko, Alexei Efros, and Trevor Darrell, "Cycada: Cycle-consistent adversarial domain adaptation," in *International conference on machine learning*. PMLR, 2018, pp. 1989–1998. 1

[3] Zak Murez, Soheil Kolouri, David Kriegman, Ravi Ramamoorthi, and Kyungnam Kim, "Image to image translation for domain adaptation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 4500–4509. 1

[4] Xun Huang, Ming-Yu Liu, Serge Belongie, and Jan Kautz, "Multimodal unsupervised image-to-image translation," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 172–189. 1

[5] Mingsheng Long, Yue Cao, Jianmin Wang, and Michael Jordan, "Learning transferable features with deep adaptation networks," in *International conference on machine learning*. PMLR, 2015, pp. 97–105. 2, 5

[6] Baochen Sun, Jiashi Feng, and Kate Saenko, "Correlation alignment for unsupervised domain adaptation," in *Domain Adaptation in Computer Vision Applications*, pp. 153–171. Springer, 2017. 2

[7] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky, "Domain-adversarial training of neural networks," *The journal of machine learning research*, vol. 17, no. 1, pp. 2096–2030, 2016. 2, 3

[8] Eric Tzeng, Judy Hoffman, Kate Saenko, and Trevor Darrell, "Adversarial discriminative domain adaptation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 7167–7176. 2

[9] Mingsheng Long, Zhangjie Cao, Jianmin Wang, and Michael I Jordan, "Conditional adversarial domain adaptation," *arXiv preprint arXiv:1705.10667*, 2017. 2

[10] Kuniaki Saito, Kohei Watanabe, Yoshitaka Ushiku, and Tatsuya Harada, "Maximum classifier discrepancy for unsupervised domain adaptation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 3723–3732. 2

[11] Yuchen Zhang, Tianle Liu, Mingsheng Long, and Michael Jordan, "Bridging theory and algorithm for domain adaptation," in *International Conference on Machine Learning*. PMLR, 2019, pp. 7404–7413. 2

[12] Shai Ben-David, John Blitzer, Koby Crammer, Alex Kulesza, Fernando Pereira, and Jennifer Wortman Vaughan, "A theory of learning from different domains," *Machine learning*, vol. 79, no. 1, pp. 151–175, 2010. 2

[13] Shuangjun Liu, Naveen Sehgal, and Sarah Ostadabbas, "Adapted human pose: Monocular 3d human pose estimation with zero real 3d pose data," *arXiv preprint arXiv:2105.10837*, 2021. 2, 5

[14] Xiheng Zhang, Yongkang Wong, Mohan S Kankanhalli, and Weidong Geng, "Unsupervised domain adaptation for 3d human pose estimation," in *Proceedings of the 27th ACM International Conference on Multimedia*, 2019, pp. 926–934. 2, 4, 5

[15] Dong-Hyun Lee et al., "Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks," in *Workshop on challenges in representation learning, ICML*, 2013, vol. 3, p. 896. 2

[16] Yang Zou, Zhiding Yu, BVK Kumar, and Jinsong Wang, "Unsupervised domain adaptation for semantic segmentation via class-balanced self-training," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 289–305. 2

[17] Yang Zou, Zhiding Yu, Xiaofeng Liu, BVK Kumar, and Jinsong Wang, "Confidence regularized self-training," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 5982–5991. 2

[18] Gül Varol, Javier Romero, Xavier Martin, Naureen Mahmood, Michael J. Black, Ivan Laptev, and Cordelia Schmid, "Learning from synthetic humans," in *CVPR*, 2017. 3

[19] 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*, 2014. 3, 4

[20] 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. 3

[21] Aiden Nibali, Zhen He, Stuart Morgan, and Luke Prendergast, "3d human pose estimation with 2d marginal heatmaps," in *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 2019, pp. 1477–1485. 4

[22] John C Gower, "Generalized procrustes analysis," *Psychometrika*, vol. 40, no. 1, pp. 33–51, 1975. 4