Domain Adaptive Pose Estimation Via Multi-level Alignment

Yugan Chen
School of computer science and engineering
Nanjing University of Science and Technology
Nanjing, China
chenyugan@njust.edu.cn

Lin Zhao*
School of computer science and engineering
Nanjing University of Science and Technology
Nanjing, China
zhaolin@njust.edu.cn

Yalong Xu
School of computer science and engineering
Nanjing University of Science and Technology
Nanjing, China
xuyalong@njust.edu.cn

Honglei Zu
School of computer science and engineering
Nanjing University of Science and Technology
Nanjing, China
zuhonglei@njust.edu.cn

Xiaoqi An
School of computer science and engineering
Nanjing University of Science and Technology
Nanjing, China
xiaoqi.an@njust.edu.cn

Guangyu Li
School of computer science and engineering
Nanjing University of Science and Technology
Nanjing, China
guangyu.li2017@njust.edu.cn

***Abstract*—Domain adaptive pose estimation aims to enable deep models trained on source domain (synthesized) datasets produce similar results on the target domain (real-world) datasets. The existing methods have made significant progress by conducting image-level or feature-level alignment. However, only aligning at a single level is not sufficient to fully bridge the domain gap and achieve excellent domain adaptive results. In this paper, we propose a multi-level domain adaptation approach, which aligns different domains at the image, feature, and pose levels. Specifically, we first utilize image style transfer to ensure that images from the source and target domains have a similar distribution. Subsequently, at the feature level, we employ adversarial training to make the features from the source and target domains preserve domain-invariant characteristics as much as possible. Finally, at the pose level, a self-supervised approach is utilized to enable the model to learn diverse knowledge, implicitly addressing the domain gap. Experimental results demonstrate that significant improvement can be achieved by the proposed multi-level alignment method in pose estimation, which outperforms previous state-of-the-art in human pose by up to 2.4% and animal pose estimation by up to 3.1% for dogs and 1.4% for sheep. The codes are available at the link.**

***Index Terms*—Unsupervised Domain Adaption, Pose Estimation, Self-supervised Learning**

I. Introduction

Recently, significant progress has been made in 2D pose estimation using deep learning methods [1]–[5]. However, these methods train high-performance models by using a large amount of labeled data, which is often time-consuming and expensive to obtain, especially with real-world datasets. To address the lack of the labels of the real-world datasets, domain adaptation (DA) [6] are being explored. By leveraging labeled virtual datasets, DA can transfer knowledge learned from virtual datasets (source domain) to unlabeled real-world datasets (target domain). Due to the development of computer technology and virtual reality techniques, labeled virtual data is much more cost-effective than real-world data. As a result, domain adaptive pose estimation has attracted a lot of attention.

While domain adaptation methods for classification tasks [7]–[10] and semantic segmentation tasks [11], [12] have seen significant progress in recent years, there has been relatively little research applying these methods to pose estimation. Existing works, whether for human pose estimation or animal pose estimation, can be broadly classified into two categories. One focuses on narrowing the domain gap between domains by performing domain alignment on the image-level, while the other primarily focuses on the pose-level. UDAPE [13] proposes a style transfer technique that aims to align domains at the image-level. RegDA [14] employs adversarial training with two different regressors to correct pose errors. CC-SSL [15] and UDA-ANIMAL [16] achieve pose-level alignment by designing different strategies to continuously update pseudo-labels. Although these methods all successfully improve the accuracy and performance of pose estimation,

This work was supported by the National Natural Science Fund of China under Grant No.62172222, the Postdoctoral Innovative Talent Support Program of China under Grant 2020M681609.

* Corresponding author.

aligning domains at a single level fails to bridge domain gap comprehensively.

In order to comprehensively bridge the domain gap, in this work, we propose a multi-level alignment framework for DA pose estimation. The framework performs image, feature and pose levels alignment to effectively bridge the gap between different domains. First, at the image-level, the source image is transferred to the style of the target image, so that they can have similar data distributions. Second, at the feature-level, we utilize adversarial learning, which is achieved by incorporating a discriminator with gradient reverse layer, to align the distribution of features between two domains. This approach effectively makes model produces domain-invariant features. Lastly, at the pose-level, we utilize an information maximization self-supervised learning technique. This enables the model to learn meaningful and diverse pose representations and prevent it from biasing toward the source domain. We evaluate the proposed method on multiple datasets and demonstrate its effectiveness in bridging the domain gap. The main contributions of our work are as follows:

- We propose a novel framework that leverages different alignment strategies at the image, feature, and pose levels to address the domain gap in cross-domain pose estimation.
- We conduct comprehensive experiments on both human and animal domain adaptive pose estimation benchmarks and achieve the state-of-the-art performance. For example, we achieve 84.4% of accuracy on the task *SURREAL* $\rightarrow$ *LSP*, which is 2.4% higher than the previous SOTA.

II. RELATED WORK

A. Pose Estimation

Pose estimation is a foundational visual task. Existing works can be mainly divided into two categories: CNN-based [1], [2], [5] and transformer-based [3], [4]. Simplebaseline [2] introduces a simple and effective model based on ResNet [17]. HRNet [1] maintains high-resolution images throughout the entire training phase, achieving significant results. ViTPose [3] achieves excellent results by utilizing a regular ViT [18] structure as the backbone and combining it with a lightweight decoder. These methods have achieved near-human-level estimation results, but they require a large amount of labeled data for training. Hence, in this work, we focus on addressing the domain gap issue in domain adaptive pose estimation, which makes it possible to train high-performance models solely using synthetic data.

B. Domain Adaptive Pose Estimation

Currently, there are two main categories of domain adaptation frameworks for pose estimation. The first category utilizes shared network structures, where the weights of the network are the same for both the source and target domains. RegDA [14] employs two independent regressors to narrow the domain gap adversarially. CC-SSL [15] utilizes consistency transformation and refinement of pseudo-labels. The second category involves non-shared network structures, typically employing a teacher-student paradigm to update weights. UDAPE [13] provides a good framework for DA pose estimation by style transfer. UDA-Animal [16] combines pseudo-label updates with the mean-teacher [19] framework. However, the aforementioned works only conduct alignment on a single level. Our work aims to achieve domain adaption through a multi-level approach.

III. METHODOLOGY

A. Preliminaries and Overview

In the source domain, a labeled pose dataset $\mathcal{S} \in \left\{\left(x_s^i, y_s^i\right)\right\}_{i=1}^{N}$ includes N samples, where $x_s^i \in \mathbb{R}^{C\times H\times W}$ represents a training image and $y_s^i \in \mathbb{R}^{K\times 2}$ represents the corresponding keypoint coordinates. Here, H, W, C and K denote the height, width, the number of channels and the number of keypoints respectively. In the target domain, an unlabeled pose dataset $\mathcal{T} \in \left\{x_t^i\right\}_{i=1}^{M}$ includes M samples. In most 2D pose estimation methods, a model is trained to output a set of heatmaps $\mathcal{H} \in \mathbb{R}^{K\times H'\times W'}$, where H' and W' are the height and width of the heatmap. The final keypoint coordinates are obtained by processing these heatmaps. Ground truth heatmaps are generated through a function (usually Gaussian) $G : \mathbb{R}^{K\times 2} \rightarrow \mathbb{R}^{K\times H'\times W'}$. The entire training process is supervised using the MSE loss by comparing the predicted heatmaps with the ground truth.

The overview of the proposed DA framework is presented in Figure 1. Our model adopts the mean-teacher [19] framework, which consists of a student model f_s and a teacher model f_t. We train our model by using three alignment methods at the image, feature and pose levels respectively. To begin with, we pretrain a pose estimation model in the labeled source domain and let both the teacher and student share the same parameters from the pretrained model initially. During the image-level alignment, we make the source domain and the target domain have similar distribution by transferring the source image style to match the target image. Furthermore, we utilize the feature adversarial learning to align the distributions of features from two domains in student model, which will allow the student model to reduce domain gap. Finally, self-supervised learning is employed at the pose-level to reduce the student's bias towards the source domain and improve the teacher's ability to generate dependable pseudo-labels.

B. Image-Level Alignment

Inspired by [13], we utilize image style transfer to align the images from two domains. Firstly, a pre-trained VGG [20] f_v is used to extract corresponding features from the content image x_s and style image x_t. Then, the generator g in AdaIN [21] is employed to generate the desired stylized output:

$$T(x_s, x_t, \eta) = g(\eta t + (1-\eta) f_v(x_s)), \tag{1}$$

where t = AdaIN$(f_v(x_s), f_v(x_t))$ and η is a weight parameter controlling the ratio between the content image and style image. Through the style transfer method, we are able to

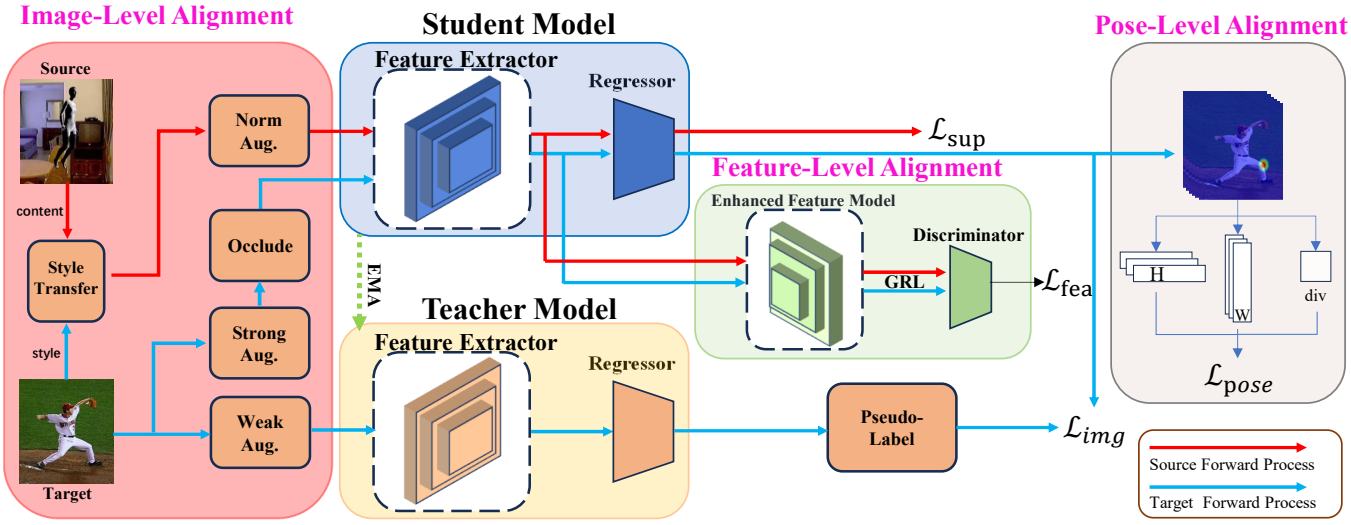

Fig. 1. **Overview of our proposed method**. The framework consists of two branches: 1) Student model for cross-domain learning and 2) Teacher model that provides pseudo-labels for the Student model. We employ image-level alignment through style transfer during the input image processing, feature-level alignment through adversarial learning and pose-level alignment through self-supervised learning to update the student model's parameters, and use exponential moving averages (EMA) to update the teacher model. **GRL** refers to Gradient Reversal Layer, which is used to align the distributions of the two domains using a discriminator with gradient inversion layers.

transform the images from the source domain into $x_{s\to t} = T(x_s, x_t, \eta)$ and narrow the domain gap at the image level.

Moreover, to prevent the model from forgetting knowledge learned from the source domain during the learning process, we still need to supervise the student model using the labels of the source domain. After style transfer, we apply normal augmentation $\mathcal{A}_1$ to the style transferred images. Therefore, the supervised loss in the student model is:

$$\mathcal{L}_{sup} = \frac{1}{N}\sum_{x_s\in\mathcal{S}} \left\| f_s(\mathcal{A}_1(x_{s\to t})) - \mathcal{H}_s \right\|_2, \tag{2}$$

where $\|\cdot\|_2$ refers to the MSE loss and $\mathcal{H}_s$ refers to the ground truth heatmap from the source domain.

To generate more reliable pseudo-labels, for the images from the target domain, we use different augmentation for the student model and teacher model. For the input of the student model, we first apply strong augmentation $\mathcal{A}_2$ to transform it, and then randomly occlude keypoints in the foreground to simulate various types of noise via an occluded operation: $\hat{x_t} = O(A_2(x_t))$. For the input of the teacher model, we only use weak augmentation $\mathcal{A}_3$. Meanwhile, we set a random threshold τ_i to simulate the occurrence of random noise on the teacher side. Only when the activation value is greater than the set threshold τ_i, do we perform consistent learning between the teacher and student models. Therefore, the loss function can be defined as follows:

$$\mathcal{L}_{img} = \sum_{x_t\in\mathcal{T}}\sum_{k=1}^{K} \mathbb{1}(h'_t \geqslant \tau_i) \left\| \tilde{\mathcal{A}}_2(h_t) - \tilde{\mathcal{A}}_3(h'_t) \right\|_2, \tag{3}$$

where $\tilde{\mathcal{A}}$ denotes the reverse operation of $\mathcal{A}$, $h_t = f_s(\hat{x_t})$ denotes the outputs of the student model, $h'_t = f_t(\mathcal{A}_2(x_t))$ denotes the normalized output heatmap of the teacher model and $\mathbb{1}(\cdot)$ denotes the function which can choose activation value.

C. Feature-Level Alignment

Considering that pseudo-labels from the teacher model are influenced by the source domain, there is a risk of unintentionally biasing the student towards that domain. In order to make the model learn domain invariant feature representation, we apply adversarial learning. Through adversarial learning, the student model can leverage information from both domains to narrow the domain gap and align distributions before applying pseudo-label supervision.

The specific method involves using the student's feature encoder F to obtain the initial features, and then using a feature enhancement model f_e to remove noise from both domains and enhance more representative information. A feature discriminator D is introduced to distinguish whether the features come from the source or the target domain. For each input sample, assume the probability of it belonging to the source domain as $D(f_e(F(\mathcal{X})))$, otherwise it is $1-D(f_e(F(\mathcal{X}))$. We use binary cross-entropy loss to update the discriminator, and the domain label d is set to 0 if the sample comes from the source domain, otherwise 1. The discriminator loss function can be represented as follows:

$$\mathcal{L}_{dis} = -d\log(D(f_e(F(\mathcal{X}))))-(1-d)\log(1-D(f_e(F(\mathcal{X})))), \tag{4}$$

where $\mathcal{X}$ denotes the inputs from source or target domain after the augmentation.

Similar to GAN [22], we need to encourage our discriminator to distinguish the features from source or target domains as accurately as possible, while also ensuring that feature encoder

can generate features that the discriminator cannot distinguish. Therefore, the final loss function can be defined as follows:

$$\mathcal{L}_{fea} = \max_{F} \min_{D} \mathcal{L}_{dis}. \tag{5}$$

In this work, we utilize the method [7] to optimize the max-min problem by generating reverse gradients between the feature encoder and discriminator.

D. Pose-Level Alignment

Due to the availability of labels only in the source domain and the student model can access inputs from both the source and target domains, the student model may exhibit a bias towards the source domain. To further narrow the gap between domains, we implicitly align domains at the pose level. Specifically, inspired by [23], an Information Maximization (IM) loss is used to achieve this goal. The IM loss aims to encourage the model to maximize the mutual information between the learned features from the input data, thereby making the model's output more deterministic at individual keypoints and more diverse on a global scale. This alignment does not involve explicit operations directly handling domain discrepancies, but it can improve the accuracy and robustness of the outputs of the student model. Therefore, it is referred to as implicit alignment.

Following the concept introduced by [24], we decompose the heatmap into H-direction and W-direction vectors using projection vectors to mitigate the impact of keypoint sparsity. Additionally, considering that projection vectors may stack some irrelevant points together, we only apply the Information Maximization to points with confidence greater than a threshold τ_p to prevent the model from incorrectly shifting to non-keypoint areas. The self-supervised loss is defined as follows:

$$\mathcal{L}_{pose} = \mathbb{E}_{x_i^t \in \mathcal{T}}(\mathcal{L}_{enth} + \mathcal{L}_{entw} + \mathcal{L}_{div}), \tag{6}$$

$$\begin{cases} \mathcal{L}_{enth} = -\sum_{k=1}^{K} \mathbb{1}(\varphi(h_t) \geqslant \tau_p) \phi(\varphi(h_t)) \log \phi(\varphi(h_t)), \\ \mathcal{L}_{entw} = -\sum_{k=1}^{K} \mathbb{1}(\psi(h_t) \geqslant \tau_p) \phi(\psi(h_t)) \log \phi(\psi(h_t)), \\ \mathcal{L}_{div} = \sum_{k=1}^{K} \phi(h_t) \log \phi(h_t), \end{cases} \tag{7}$$

where $\varphi(\cdot)$ denotes the operation to map $\cdot$ to a vector in direction of H, $\psi(\cdot)$ denotes the operation to map $\cdot$ to a vector in direction of W, and $\phi(\cdot)$ denotes the operation of softmax.

E. Training Details

In summary, the entire objective loss function can be defined as:

$$\mathcal{L} = \mathcal{L}_{sup} + \alpha \mathcal{L}_{img} + \beta \mathcal{L}_{fea} + \gamma \mathcal{L}_{pose}, \tag{8}$$

where α, β and γ are trade-off parameters.

After conducting all the alignments, the student model can undergo normal gradient updates, while the parameters of the teacher are updated using the parameters of the student model through exponential moving average (EMA):

$$\theta_t \leftarrow \mu \theta_t + (1 - \mu) \theta_s, \tag{9}$$

where θ_t and θ_s denote the model parameters of the teacher and student, respectively. μ denotes the smoothing coefficient and the out default setting is 0.999.

IV. Experiments

To validate the effectiveness of our approach on pose estimation tasks, we conduct experiments on benchmark datasets including human and animal pose estimation and compare our method with previous state-of-the-art models.

Datasets. For human pose estimation, *SURREAL* [25] and *Leeds Sports Pose* (LSP) [26] datasets are used. *SURREAL* is a synthetic dataset used as the source domain, comprising a total of 6 million images. *Leeds Sports Pose* is a real-world dataset containing 2k annotated images of athletes' poses. Our task is to perform domain adaptation from *SURREAL* as the source domain to *LSP* as the target domain.

We utilized three datasets for animal pose estimation. The first one is *SynAnimal* [15], which is a synthetic dataset generated by rendering CAD models. It consists of five different animals: horse, tiger, sheep, hound and elephant. Each animal category contains 10k images. The second dataset is *TigDog* [27], a real-world dataset obtained by slicing frames from videos. It includes 30k images of horses and tigers. The third dataset is *AnimalPose* [28], which contains 6.1k real-world animal images of dogs, cats, cows, sheep, and horses. Our task is to perform domain adaptation from *SynAnimal* as the source domain to *TigDog* and *AnimalPose* as the target domains.

Implementation Details. The pose estimation network is based on SimpleBaseline [2]. We use a pre-trained ResNet101 as the backbone network. The Adam [29] optimizer is employed with an initial learning rate of 1e-4, which gradually decreases to 1e-5 at the 22500th iteration and finally reaches 1e-6 after 30000 iterations. As for hyperparameters, we set α = 1, β = 0.1 and σ = 0.3. After the training is completed, the teacher model is used for the final inference.

A. Main Results

In this section, the baselines, metrics and quantitative results on different tasks are shown.

Baselines. We compare our approach with the following baseline methods in domain adaptation for pose estimation: CC-SSL [15], RegDA [14], and UDAPE [13]. All baseline methods utilize ResNet-101 as the backbone.

Metrics. To evaluate our approach, the Percentage of Correct Keypoints (PCK) is used as the metric, which gives the percentage of correctly estimated keypoints. In Tables 1-5, we report PCK@0.05 that measures the percentage of accurate predictions, with a threshold of 5% relative to the image size. For the human pose estimation task, we follow the same setting as other methods [13], [14] and focus on the following keypoints: Shoulder (Sld), Elbow (Elb), Wrist, Hip, Knee, and Ankle. As for the animal pose estimation tasks, we also follow

TABLE I
PCK@0.05 ON TASK *SURREAL → LSP*

method	Sld	Elb	Wrist	Hip	Knee	Ankle	All
Source-only	50.6	64.8	63.3	70.1	71.2	70.1	65.0
CCSSL [15] (CVPR'20)	36.8	66.3	63.9	59.6	67.3	70.4	60.7
UDA-Animal [16](CVPR'21)	61.4	77.7	75.5	65.8	76.7	78.3	69.2
RegDA [14] (CVPR'21)	62.7	76.7	71.1	81.0	80.3	75.3	74.6
UDAPE [13] (ECCV'22)	69.2	84.9	83.3	85.5	84.7	84.4	82.0
Ours	**78.2**	**86.6**	**83.7**	**87.1**	**85.2**	**85.5**	**84.4**

the common settings as other methods [13], [16]. In *TigDog*, we select the Eye, Chin, Shoulder (Sld), Hip, Elbow (Elb), Knee, and Hoof. In *AnimalPose*, the Eye, Hoof, Knee, and Elbow are selected.

Results on Human Pose Estimation. Table 1 presents the results of the human pose estimation task *SURREAL → LSP*. Our method achieves the best performance outperforming the previous state-of-the-art method UDAPE [13] by 2.4%. Moreover, our method consistently surpasses the baseline methods in every keypoint for human pose estimation.

Results on Animal Pose Estimation. Tables 2-5 present the quantitative comparing results of animal pose estimation tasks: *SynAnimal → TigDog* and *SynAnimal → AnimalPose*. On *TigDog*, our method achieves the best results on the tiger category (Table 2), surpassing UDAPE [13] by 0.5%. On the horse category (Table 3), the performance of our method is on par with UDAPE and slightly lower than that of UDA-Animal. On *AnimalPose*, our model achieves the best results on the dog and sheep categories, improving the previous best results by 3.1% and 1.4%, respectively.

TABLE II
PCK@0.05 ON TASK *SynAnimal → TigDog (Tiger)*

method	Eye	Chin	Sld	Hip	Elb	Knee	Hoof	All
Source-only	85.4	81.8	44.6	70.8	39.6	48.4	55.5	54.8
CCSSL [15] (CVPR'20)	94.3	91.3	49.5	70.2	**53.9**	59.1	70.2	66.7
RegDA [14] (CVPR'21)	93.3	92.8	50.3	67.8	50.2	55.4	60.7	61.8
UDA-Animal [16](CVPR'21)	98.4	87.2	49.4	**74.9**	49.8	62.0	73.4	67.7
UDAPE [13] (ECCV'22)	**98.5**	**96.9**	56.2	63.7	52.3	62.8	72.8	67.9
Ours	98.0	95.4	**60.4**	64.1	52.1	**63.3**	**73.6**	**68.4**

TABLE III
PCK@0.05 ON TASK *SynAnimal → TigDog (Horse)*

method	Eye	Chin	Sld	Hip	Elb	Knee	Hoof	All
Source-only	82.0	90.0	59.2	79.5	65.8	66.9	57.7	67.4
CCSSL [15] (CVPR'20)	89.3	92.6	69.5	78.1	70	73.1	65	73.1
RegDA [14] (CVPR'21)	89.2	92.3	70.5	77.5	71.5	72.7	63.2	73.2
UDA-Animal [16](CVPR'21)	86.9	**93.7**	**76.4**	**81.9**	70.6	**79.1**	**72.6**	**77.5**
UDAPE [13] (ECCV'22)	**91.3**	92.5	74.0	74.2	75.8	77.0	66.6	76.4
Ours	82.3	92.7	72.6	69.2	**76.4**	77.8	68.5	76.3

TABLE IV
PCK@0.05 ON TASK *SynAnimal → AnimalPose (Dog)*

method	Eye	Hoof	Knee	Elb	All
Source-only	38.2	43.2	25.7	24.1	32.0
CCSSL [15] (CVPR'20)	34.7	37.4	25.4	19.6	27.0
RegDA [14] (CVPR'21)	46.8	54.6	32.9	31.2	40.6
UDA-Animal [16](CVPR'21)	26.2	39.8	31.6	24.7	31.1
UDAPE [13] (ECCV'22)	56.1	**59.2**	38.9	32.7	45.4
Ours	**70.6**	59.1	**40.2**	**35.2**	**48.5**

TABLE V
PCK@0.05 ON TASK *SynAnimal → AnimalPose (Sheep)*

method	Eye	Hoof	Knee	Elb	All
Source-only	59.9	60.7	46.2	31.0	47.9
CCSSL [15] (CVPR'20)	44.3	55.4	43.5	28.5	42.8
RegDA [14] (CVPR'21)	62.8	68.5	57.0	42.4	56.9
UDA-Animal [16](CVPR'21)	48.2	52.9	49.9	29.7	44.9
UDAPE [13] (ECCV'22)	61.6	**77.4**	57.7	44.6	60.2
Ours	**66.8**	75.8	**61.6**	**44.8**	**61.6**

B. Ablation Study

We conduct ablation studies to investigate the effectiveness of each module proposed in our method.

Effect of loss functions. We evaluate the performance of each component of our framework through the task *SURREAL → LSP* in Table 6. We adopt the mean-teacher method [19] as our baseline. Building upon image-level alignment, we introduce additional loss functions for both feature-level alignment and pose-level alignment: adversarial loss and self-supervised loss respectively. Experimental results demonstrate that feature-level alignment significantly enhances the model's performance, resulting in a 0.7% improvement. Pose-level alignment alone does not provide significant benefits, yielding a marginal improvement of 0.3%. Nevertheless, when both pose-level and feature-level alignment are conducted, an additional 1.1% improvement is observed. In summary, our proposed multi-level alignment strategy proves to be crucial in bridging the domain gap.

TABLE VI
THE EFFECT OF LOSS FUNCTIONS ON TASK *SURREAL → LSP*

method	Sld	Elb	Wrist	Hip	Knee	Ankle	All
MT	57.1	70.6	65.7	74.6	76.1	74.5	69.8
$\mathcal{L}_{img}$	69.3	86.3	84.5	86.3	84.6	84.4	82.6
$\mathcal{L}_{img}$&$\mathcal{L}_{fea}$	72.8	86.0	**84.9**	85.9	85.0	85.0	83.3
$\mathcal{L}_{img}$&$\mathcal{L}_{pose}$	71.2	86.4	84.5	86.4	84.7	84.5	82.9
$\mathcal{L}_{img}$&$\mathcal{L}_{fea}$&$\mathcal{L}_{pose}$	**78.2**	**86.6**	83.7	**87.1**	**85.2**	**85.5**	**84.4**

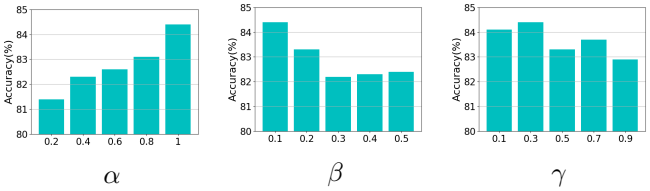

Fig. 2. Analysis of the influences of parameters on *SURREAL → LSP*.

Analysis of parameters. We illustrate the sensitivity of the parameters α, β and γ in Equation 8, through the task *SURREAL → LSP*. The results are shown in Fig 2. Increasing α shows better performance and confirms the effectiveness of pre-alignment. However, increasing β leads to a decline in performance and potentially causes overfitting. Lastly, the performance for γ remains stable with little variation. It can be observed that our model exhibits the best performance when $\alpha = 1$, $\beta = 0.1$ and $\gamma = 0.3$.

V. Conclusion

In this paper, we propose a novel approach for unsupervised domain adaption on pose estimation. Based on the mean-teacher framework, we propose multi-level domain alignment strategy including image-level alignment through style transfer, feature-level alignment through adversarial learning and pose-level alignment through self-supervised learning method. The strategy effectively alleviates the domain gap and the source domain bias issues. The effectiveness and superiority of this approach have been verified in human and animal pose estimation domain adaption tasks.

References

[1] Ke Sun, Bin Xiao, and Dong Liu, "Deep high-resolution representation learning for human pose estimation," in *CVPR*, 2019, pp. 5693–5703.

[2] Bin Xiao, Haiping Wu, and Yichen Wei, "Simple baselines for human pose estimation and tracking," in *ECCV*, 2018, pp. 466–481.

[3] Yufei Xu, Jing Zhang, and Qiming Zhang, "Vitpose: Simple vision transformer baselines for human pose estimation," in *NeurIPS*, 2022, pp. 38571–38584.

[4] Xinyu Yi, Yuxiao Zhou, and Feng Xu, "Transpose: Real-time 3d human translation and pose estimation with six inertial sensors," *TOG*, vol. 40, no. 4, pp. 1–13, 2021.

[5] Lin Zhao, Nannan Wang, and Chen Gong, "Estimating human pose efficiently by parallel pyramid networks," *TIP*, vol. 30, pp. 6785–6800, 2021.

[6] Kate Saenko, Brian Kulis, Mario Fritz, and Trevor Darrell, "Adapting visual category models to new domains," in *ECCV*, 2010, pp. 213–226.

[7] Yaroslav Ganin and Victor Lempitsky, "Unsupervised domain adaptation by backpropagation," in *ICML*, 2015, pp. 1180–1189.

[8] Judy Hoffman, Dequan Wang, and Fisher Yu, "Fcns in the wild: Pixel-level adversarial and constraint-based adaptation," *arXiv preprint arXiv:1612.02649*, 2016.

[9] Mingsheng Long, Yue Cao, and Jianmin Wang, "Learning transferable features with deep adaptation networks," in *ICML*, 2015, pp. 97–105.

[10] Yuchen Zhang, Tianle Liu, and Mingsheng Long, "Bridging theory and algorithm for domain adaptation," in *ICML*, 2019, pp. 7404–7413.

[11] Lukas Hoyer, Dengxin Dai, Haoran Wang, and Luc Van Gool, "Mic: Masked image consistency for context-enhanced domain adaptation," in *CVPR*, 2023, pp. 11721–11732.

[12] Yunsheng Li, Lu Yuan, and Nuno Vasconcelos, "Bidirectional learning for domain adaptation of semantic segmentation," in *CVPR*, 2019, pp. 6936–6945.

[13] Donghyun Kim, Kaihong Wang, and Kate Saenko, "A unified framework for domain adaptive pose estimation," in *ECCV*, 2022, pp. 603–620.

[14] Junguang Jiang, Yifei Ji, and Ximei Wang, "Regressive domain adaptation for unsupervised keypoint detection," in *CVPR*, 2021, pp. 6780–6789.

[15] Jiteng Mu, Weichao Qiu, and Gregory D Hager, "Learning from synthetic animals," in *CVPR*, 2020, pp. 12386–12395.

[16] Chen Li and Gim Hee Lee, "From synthetic to real: Unsupervised domain adaptation for animal pose estimation," in *CVPR*, 2021, pp. 1482–1491.

[17] Kaiming He, Xiangyu Zhang, and Shaoqing Ren, "Deep residual learning for image recognition," in *CVPR*, 2016, pp. 770–778.

[18] Alexey Dosovitskiy, Lucas Beyer, and Alexander Kolesnikov, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[19] Antti Tarvainen and Harri Valpola, "Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results," in *NeurIPS*, 2017.

[20] Karen Simonyan and Andrew Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[21] Xun Huang and Serge Belongie, "Arbitrary style transfer in real-time with adaptive instance normalization," in *ICCV*, 2017, pp. 1501–1510.

[22] Ian Goodfellow, Jean Pouget-Abadie, and Mehdi Mirza, "Generative adversarial nets," in *NeurIPS*, 2014.

[23] Jian Liang, Dapeng Hu, and Jiashi Feng, "Do we really need to access the source data? source hypothesis transfer for unsupervised domain adaptation," in *ICCV*, 2020, pp. 6028–6039.

[24] Qucheng Peng, Ce Zheng, and Chen Chen, "Source-free domain adaptive human pose estimation," in *ICCV*, 2023, pp. 4826–4836.

[25] Gul Varol, Javier Romero, and Xavier Martin, "Learning from synthetic humans," in *CVPR*, 2017, pp. 109–117.

[26] Sam Johnson and Mark Everingham, "Clustered pose and nonlinear appearance models for human pose estimation.," in *BMVC*, 2010, p. 5.

[27] Luca Del Pero, Susanna Ricco, and Rahul Sukthankar, "Articulated motion discovery using pairs of trajectories," in *CVPR*, 2015, pp. 2151–2160.

[28] Jinkun Cao, Hongyang Tang, and Hao-Shu Fang, "Cross-domain adaptation for animal pose estimation," in *ICCV*, 2019, pp. 9498–9507.

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