Source-free Domain Adaptive Human Pose Estimation

Qucheng Peng, Ce Zheng, Chen Chen

Center for Research in Computer Vision, University of Central Florida

{qucheng.peng,ce.zheng}@ucf.edu, chen.chen@crcv.ucf.edu

Abstract

Human Pose Estimation (HPE) is widely used in various fields, including motion analysis, healthcare, and virtual reality. However, the great expenses of labeled real-world datasets present a significant challenge for HPE. To overcome this, one approach is to train HPE models on synthetic datasets and then perform domain adaptation (DA) on real-world data. Unfortunately, existing DA methods for HPE neglect data privacy and security by using both source and target data in the adaptation process.

To this end, we propose a new task, named source-free domain adaptive HPE, which aims to address the challenges of cross-domain learning of HPE without access to source data during the adaptation process. We further propose a novel framework that consists of three models: source model, intermediate model, and target model, which explores the task from both source-protect and target-relevant perspectives. The source-protect module preserves source information more effectively while resisting noise, and the target-relevant module reduces the sparsity of spatial representations by building a novel spatial probability space, and pose-specific contrastive learning and information maximization are proposed on the basis of this space. Comprehensive experiments on several domain adaptive HPE benchmarks show that the proposed method outperforms existing approaches by a considerable margin. The codes are available at https://github.com/davidpengucf/SFDAHPE.

1. Introduction

The accuracy of 2D human pose estimation (HPE) from monocular images has been significantly improved through the use of deep neural networks. However, obtaining sufficient labeled data for training such models is a laborious and time-consuming task, particularly for real-world datasets that require diverse appearances and poses with ground truth annotations. In contrast, the development of computer graphics and game engines has facilitated the generation of annotated synthetic datasets that can generate a virtually unlimited number of labeled samples. Nevertheless, models trained on synthetic data may not perform well on real-world data due to the domain gap or differences in data distribution. To address this challenge, domain adaptation (DA) is essential to enable synthetic data-trained models to generalize to real-world data.

Recently, source-free domain adaptation has gained significant attention due to its practicality in real-world scenarios, where using labeled source data during the adaptation is expensive or not feasible due to data privacy and security concerns. However, existing source-free domain adaptation methods in classification tasks may not be directly applicable to more complex tasks, such as 2D human pose estimation (HPE). This is because HPE requires estimating the spatial locations of different body joints, instead of just the object's category as classification does, making it more challenging to adapt from source to target domains. Therefore, there is a pressing need to develop effective source-free domain adaptation techniques specifically tailored to HPE tasks. This motivates us to propose a new task named source-free domain adaptive HPE in this paper.

The task of source-free domain adaptive HPE presents challenges from *both the source-free DA and pose estimation perspectives*. Specifically, in traditional DA tasks, source data can participate in the adaptation process, allowing for a focus on reducing domain shifts between source and target. However, in source-free DA, only the pretrained source model can be used during adaptation, leading to catastrophic forgetting of the source domain. Directly applying the source pretrained model can also introduce noise due to domain shifts. Therefore, balancing the absorption of the source model's knowledge with the resistance of the source model's noise is critical. Current methods tend to adopt a progressive replacement strategy to alleviate the runoff of source information. For example, [21, 22] use the previous model to supervise ongoing adaptation, thereby evoking the source-related memory, while [32, 6] use the exponential moving average (EMA) strategy to slow down the forgetting of the source. Although these methods alleviate the negative impact of noise successfully, they modify the overall parameters of the pretrained source model

significantly, leading to the disappearance of source-related knowledge and representations during the adaptation process. Therefore, it is necessary to develop source-protect modules that can protect source information more effectively while resisting noise at the same time.

Moreover, we outline another challenge of source-free domain adaptive HPE: the sparsity of keypoints (i.e. key body joints) in an image. When it comes to domain adaptation (DA) in classification, achieving distribution alignment to reduce domain gap is a straightforward process as the outputs can be represented in a probability space that includes all the categories. However, the same cannot be said for Human Pose Estimation (HPE) based on a complete image, as each keypoint has a specific distance from others based on bone lengths, and the number of keypoints K (e.g., 14 for human pose and 21 for hand pose) is significantly smaller than the number of pixels in an image (e.g., $512 \times 512 \approx 0.26M$), making it a challenging task. Although heatmaps [34, 33] are commonly used to predict keypoints (heatmap representation-based 2D HPE is considered as the mainstream approach for HPE) while reducing dimensionality in HPE, the number of pixels inside the heatmap is still considerable (e.g. $128 \times 128 \approx 16K$) in comparison to K. Therefore, building a spatial probability space for the keypoints based on either the entire image or the heatmap is not ideal. [12] proposes constructing a low-dimensional distribution based solely on the K keypoints, effectively transforming the HPE task into a K-classification task. However, this spatial distribution relies on clear decision boundaries between distinctive keypoints, which is not reasonable in HPE. In other words, all samples can be classified to a certain category in classification, but most pixels inside an image (e.g. more than 97% if the fault tolerance is decided by the metric PCK@0.05) can not be assigned to a certain keypoint in HPE, but mispredictions may occur at these unpresented locations. Hence, it is necessary to develop a novel spatial probability space that reduces sparsity while providing complete representations of the prediction space.

To address the aforementioned task-specific challenges, we propose a novel framework that incorporates source-free domain adaptive HPE from both source-protect and target-relevant perspectives. Our framework consists of three models: the source model, the intermediate model, and the target model. Source model preserves source information, and target model absorbs knowledge from target data, while intermediate model interacts with both of them to reduce the domain gap between source and target implicitly, thereby used for the final inference. In the source-protect modules, we conduct knowledge transfer between the source model and the intermediate model to prevent catastrophic forgetting of the source-related information. We transfer knowledge from the source model to the intermediate model while keeping the source model's feature extractor fixed. To re-

sist noise, we propose a new residual loss and finetune the source model's regressor. In the target-relevant modules, we conduct executions between the intermediate model and the target model to mitigate domain shifts between the source and target domains. We build a new spatial probability space for HPE to alleviate the sparsity problem. On the basis of this space, we propose contrastive learning and information maximization designed specifically for human pose estimation. Our contributions are summarized as follows:

- We construct a new task named source-free domain adaptive human pose estimation, which focuses on the cross-domain learning of HPE without access to source data during the adaptation process.
- We propose a new framework consisting of three models: source model, intermediate model, and target model, which explores the task from both source-protect and target-relevant perspectives.
- We conduct comprehensive experiments on several domain adaptive HPE benchmarks and the results show that our method outperforms the existing approaches by a considerable margin.

2. Related Work

Human Pose Estimation. The heatmap-based method [29, 37] is currently a widely used technique in 2D HPE, which involves generating heatmaps by placing 2D Gaussian kernels on potential keypoints, followed by converting these heatmaps to coordinates. For instance, [3] uses global- and refine-nets, while [39] simplifies the network structures through the Simple Baselines technique. Some approaches have focused on utilizing features at different scales [30], while others aim to improve the quality of heatmaps by extending Gaussians to second order [40] or using coupled embeddings for heatmap regression [37]. *In this study, we adopt the same heatmap-based HPE methods as other domain adaptive HPE works [12, 16].*

General Domain Adaptation. There are two main categories of methods for achieving general domain adaptation, which requires access to both the source and target data during adaptation. The first category is metric-based methods that explicitly measure the discrepancy between the source and target domains. For example, [24] uses maximum mean discrepancy, and [35] applies deep domain confusion. Recent work, such as [4], jointly performs clustering and discrimination for alignment using contrastive learning [2, 13]. The second category is GAN-based [7] methods that build a min-max game for two players related to the source and target domains. For example, [5, 28] adopts domain confusion to confuse the two players, while [31, 27] shows that discriminative clustering on the target benefits adaptation. *However, access to source data is required in the general domain adaptation setting, which raises concerns about data privacy and security.*

Source-Free Domain Adaptation. Source-Free has become the mainstream paradigm for alleviating concerns about data privacy and security in domain adaptation, as shown in the work of [20, 21, 26]. There are two technique routes under the source-free setting: self-supervision and virtual source transfer. For the self-supervised methods, [21] is the most representative, which introduces information maximization to assist adaptation. Other self-supervised methods include [38], which treats the problem from a discriminative perspective, and [1, 26], which proposes online pseudo label refinement. In virtual source methods, most build GANs to generate virtual source data. Examples include conditional GANs used in [18], and collaborative GANs in [20], which achieve better generations. *However, most source-free DA methods for classification are not applicable to HPE tasks due to the sparsity of keypoints' distributions.*

Domain Adaptive Human Pose Estimation. There are two main categories of methods used for domain-adaptive HPE. The first category involves a shared structure, where the weights of the networks from both the source pretrain and target adaptation are shared. CC-SSL [25] employs a single end-to-end trained network, RegDA [12] utilizes one shared feature extractor and two separate regressors, and TransPar [8] emphasizes transferable parameters using a similar structure. The second category is the unshared structure, where the pretrain and adaptation create a teacher-student paradigm. In this category, MDAM [19] addresses this structure alongside a novel pseudo-label strategy, UniFrame [16] modifies the classic Mean-Teacher [32] model by combining it with style transfer [10], and MarsDA [14] applies a teacher-student approach to edit RegDA. *However, **all these existing domain adaptive HPE approaches are not source-free**, which raises concerns about data privacy and security. Therefore, it is necessary to consider source-free domain adaptive HPE. To our knowledge, we are the first work to study this challenging problem.*

3. Method

3.1. Preliminary of 2D HPE

In 2D HPE, we are given a labeled pose dataset $\mathcal{D} = (x_i, y_i)$, where $x_i \in \mathbb{R}^{C \times H \times W}$ is an image and $y_i \in \mathbb{R}^{K \times 2}$ are the corresponding keypoint coordinates. Here, H and W represent the height and width of the image, respectively, and C is the number of channels. Additionally, K represents the number of keypoints. To ensure smoother training, most existing 2D methods represent the keypoint coordinates as heatmaps $HM_i \in \mathbb{R}^{K \times H' \times W'}$, where H' and W' are the height and width of the heatmaps, respectively. These heatmaps are used to represent the coordinates in a spatial way. The process of how the model works is defined
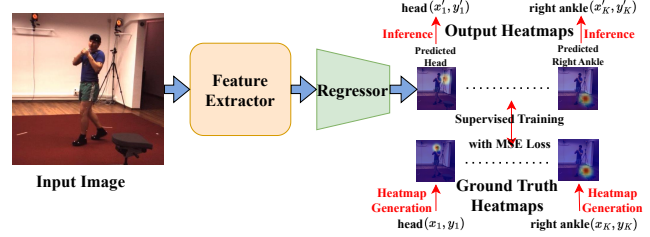

Figure 1. Process of heatmap-based 2D HPE. The model includes two components: the feature extractor G and the regressor F. After feeding the input image to the model, we obtain K output heatmaps corresponding to K keypoints.

in Fig. 1. Ground truth heatmaps are generated by a 2D Gaussian centered at the ground truth joint location, with one heatmap for each keypoint. We denote the transform from coordinates to heatmaps as $T : \mathbb{R}^{K \times 2} \to \mathbb{R}^{K \times H' \times W'}$. The training processes are conducted between the output heatmaps and ground truth heatmaps using MSE for supervised learning. In the inference stage, well-trained supervised models output heatmaps that predict the probability of joints occurring at each pixel. For each heatmap, the pixel with the highest probability is considered to be the predicted location of the joint.

3.2. Problem Statement

In the source-free domain adaptive HPE setting, we have a source domain dataset $\mathcal{S} = \{(x_i^s, y_i^s)\}_{i=1}^{n_s}$ with n_s labeled pose samples. $x_i^s \in \mathbb{R}^{C \times H \times W}$ is the source image and $y_i^s \in \mathbb{R}^{K \times 2}$ is the corresponding pose annotation. Besides, there exists a target domain dataset $\mathcal{T} = \{x_i^t\}_{i=1}^{n_t}$ that includes n_t unlabeled pose samples. The source and target domains share the same label space but lie in different distributions (e.g. synthetic vs. real). The training procedure is split into two stages: the pretrain stage and the adaptation stage. In the pretrain stage, the labeled source domain dataset $\mathcal{S}$ is used to train the source model f^{sr}. However, what distinguishes source-free domain adaptive HPE from general domain adaptive HPE is the adaptation stage, where only the source model f^{sr} and target domain dataset $\mathcal{T}$ are available for training, and the source domain dataset cannot be used anymore. Our objective is to develop a model that performs well on the target dataset.

3.3. Pipeline of Our Method

3.3.1 Source Model Pretrain

For DA in HPE, labeled source data are applied for pretraining. Same as most existing domain adaptive pose estimation methods [25, 19, 12, 16], the network is composed of a feature extractor G (such as ResNet backbone) and a pose regressor F, so the source model can be represented as $f^{sr} = F^{sr}(G^{sr}(\cdot))$. Based on the supervised heatmap-based loss (MSE loss), we propose the overall objective for
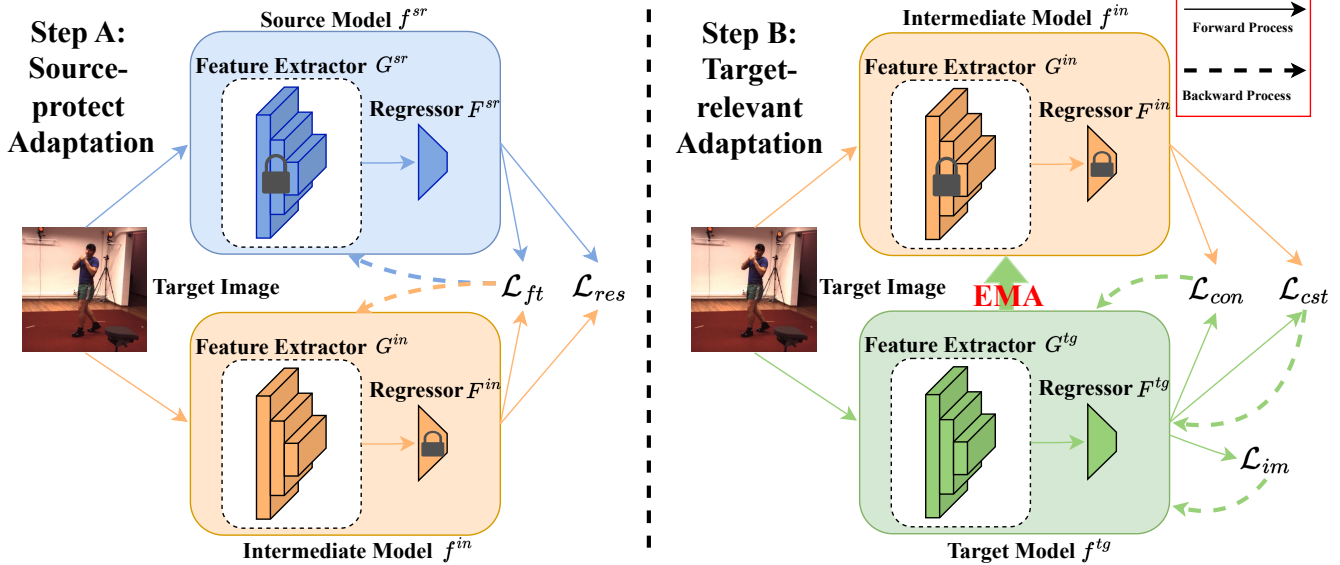
Figure 2. Overall framework of our method. It includes three models, each with one feature extractor and one regressor. The adaptation contains two steps: Step A and Step B. Step A is applied for source-protect adaptation, while Step B aims at target-relevant adaptation. The solid arrows show the forward process of associated losses, and the dotted arrows show the backward updating process of associated losses. EMA means the exponential moving average process from target model to intermediate model to update intermediate model's weights.

source pretraining:

$$\mathcal{L}_{pretrain} = \mathbb{E}_{(x_i^s, y_i^s) \in \mathcal{S}} \mathcal{L}_{mse}(F^{sr}(G^{sr}(x_i^s)), T(y_i^s)), \quad (1)$$

where $\mathcal{L}_{mse}$ is the Mean Squared Error (MSE) loss between the heatmap representations of prediction and ground truth, and T is the transform from coordinates to heatmaps defined in Sec. 3.1.

3.3.2 Adaptation Framework

As Fig. 2 shows, our method contains three models: source model $f^{sr} = F^{sr}(G^{sr}(\cdot))$, intermediate model $f^{in} = F^{in}(G^{in}(\cdot))$, and target model $f^{tg} = F^{tg}(G^{tg}(\cdot))$. Each model includes one feature extractor and one regressor. Assume the heatmaps generated by source model as $\boldsymbol{H_i^{sr}} = F^{sr}(G^{sr}(x_i^t)) = \left[h_{i,1}^{sr}, h_{i,2}^{sr}, ..., h_{i,K}^{sr} \right]^\top$, and those from intermediate model and target model respectively as $\boldsymbol{H_i^{in}} = F^{in}(G^{in}(x_i^t)) = \left[h_{i,1}^{in}, h_{i,2}^{in}, ..., h_{i,K}^{in} \right]^\top$ and $\boldsymbol{H_i^{tg}} = F^{tg}(G^{tg}(x_i^t)) = \left[h_{i,1}^{tg}, h_{i,2}^{tg}, ..., h_{i,K}^{tg} \right]^\top$, where K is the number of keypoints. We also assume the keypoint indexing space as $\mathcal{P} = [1, 2, ..., K]$. The source model is used to maintain source information, while the target model is applied to learn target knowledge. The intermediate model is the transition model between these two models. It serves as a bridge between source and target as it learns from source and target simultaneously, thus eliminating the domain gap implicitly. Therefore, it outperforms both the source and target models and is used for the final inference. When starting the adaptation, f^{in} and f^{tg} are initialized from the source model f^{sr}.

For a single adaptation iteration, two steps are executed as shown in Fig. 2. Step A is for the source-protect adaptation, and Step B aims to adapt in a target-relevant way.

In Step A, our goal is to overcome catastrophic forgetting on source domain, while eliminating the noise from the source model due to domain shift. Therefore, the source model's feature extractor G^{sr} is fixed to retain source knowledge, while its regressor F^{sr} keeps updating to improve regression accuracy, hence reducing noises in the outputs. On the opposite, the intermediate model's regressor F^{in} is fixed, but its feature extractor G^{in} continues to update. That's because we want the feature extractor to learn more source knowledge on the representation level, while the noise from the source model due to domain shifts can be inhibited.

In Step B, we aim to conduct the adaptation between the intermediate model and the target model in a target-relevant manner. During the process, several pose-specific techniques are utilized to minimize the discrepancy between these two models. Though both f^{in} and f^{tg} are updated at this step, f^{tg} is updated via back-propagation while f^{in} is via exponential moving average from f^{tg}, since averaging model weights over training steps tends to produce a more accurate model than using the final weights directly [32].

In the following two subsections, we introduce the detailed techniques used for Step A and Step B.
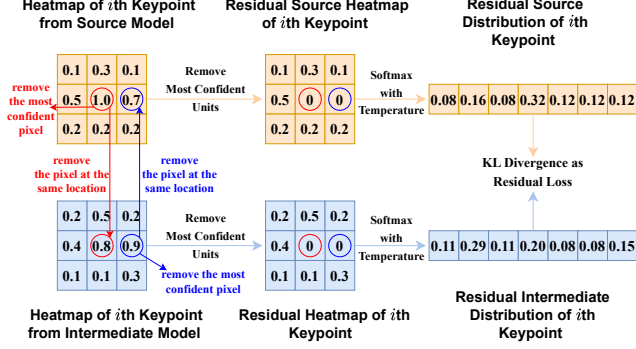
Figure 3. Proposed residual loss. Here we choose two heatmaps at the same location generated from source model and intermediate model for illustration. For simplicity, here the heatmap's scale is set as 3×3, while the actual size is 64×64.

3.3.3 Source-protect Modules

In this part, we propose modules related to source-free-specific adaptation. The biggest challenge is to distill knowledge from the source model and resist the potential noise due to domain shift. First, we introduce the finetune loss that calibrates on the source model:

$$\mathcal{L}_{ft} = \mathbb{E}_{x_i^t \in \mathcal{T}} \mathcal{L}_{mse}(F^{sr}(G^{sr}(x_i^t)), F^{in}(G^{in}(x_i^t))). \quad (2)$$

To be notified, we hope the regression results of the source model can be more accurate, while the source knowledge can be retained. Thus, the feature extractor is fixed so that the source representations continue to be generated, and the regressor is being updated:

$$\theta_{F^{sr}}^{t+1} = \theta_{F^{sr}}^t - \lambda^{sr} \frac{\partial L_{ft}}{\partial \theta_{F^{sr}}^t}, \quad (3)$$

where θ_F^s is the weights of the source model's regressor F^s, while t denotes the training step and λ^s is the source model's learning rate. However, improving source model is not enough. What's more, we propose the residual loss to improve intermediate model.

The application of Equation 2 facilitates mutual knowledge transfer between the source and intermediate models. Apart from the pixel with the highest confidence in a heatmap, the other pixels also contain valuable information that benefits the knowledge transfer, and that's why we propose residual loss in Fig. 3. From the example in the figure, the most confident pixel (confidence = 1.0) in the source heatmap is removed, along with the pixel at the same location which has a confidence of 0.8 in the intermediate heatmap marked with red. The pair marked with blue follows the same rule but on the basis of intermediate heatmap. Then we build adaptive residual heatmaps denoted as $\hat{h}_{i,j}^{sr}$ and $\hat{h}_{i,j}^{in}$. Moreover, we employ KL divergence to ensure that $\hat{h}_{i,j}^{in}$ approaches $\hat{h}_{i,j}^{sr}$, thereby preserving the source in-
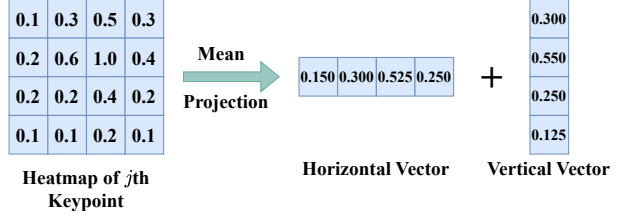

Figure 4. Illustration of the projection of a heatmap. In order to reduce the sparsity of a heatmap, we project it into a horizontal vector and a vertical vector. For simplicity, here the heatmap's scale is set as 4×4 and the vector's scale as 4×1, while the actual heatmap's size is 64×64 and the vector's size is 64×1.

formation. Based on these, we deduce the residual loss as:

$$\mathcal{L}_{res} = \mathbb{E}_{x_i^t \in \mathcal{T}} \mathbb{E}_{j \in \mathcal{P}} D_{KL}(\sigma(\hat{h}_{i,j}^{sr}/\tau)||\sigma(\hat{h}_{i,j}^{in}/\tau)), \quad (4)$$

where D_{KL} is the KL divergence, and $\sigma(\cdot)$ is the softmax function. Besides, τ is the temperature used to scale the residual heatmaps and is empirically set to 0.3. By combining the finetune loss and the residual loss, we get the objective function with the optimization process for updating the intermediate model:

$$\mathcal{L}_{in} = \mathcal{L}_{ft} + \alpha \mathcal{L}_{res}, \quad \theta_{G^{in}}^{t+1} = \theta_{G^{in}}^t - \lambda^{in} \frac{\partial L_{in}}{\partial \theta_{G^{in}}^t}, \quad (5)$$

where t is the training step, and λ^{in} is the learning rate for G^{in}. Moreover, α is the trade-off hyperparameter between these two losses.

3.3.4 Target-relevant Modules

This section introduces a module that is associated with target information. As previously mentioned, the sparsity of keypoints in the image presents a significant challenge in our task. To address this issue, we propose using projected horizontal and vertical vectors of heatmaps as spatial representations, as shown in Fig. 4. This approach reduces the sparsity of the spatial probability space effectively since the number of overall digits inside the heatmap is $64 \times 64 \approx 4K$, while the number of units in each vector is 64. Therefore, we can effectively reduce the sparsity of the spatial probability space. With 64 units in each vector, the spatial probability space size is acceptable compared to the number of keypoints.

Despite the reduction in sparsity, the completeness of spatial representations is not affected. That is because all the coordinates can always be divided into two representations, horizontally and vertically. In such a case, our spatial representation is better than the K-classification probability space proposed by [12], which only includes the keypoint-relevant information and does not take the units in the heatmap that are not keypoints into consideration.
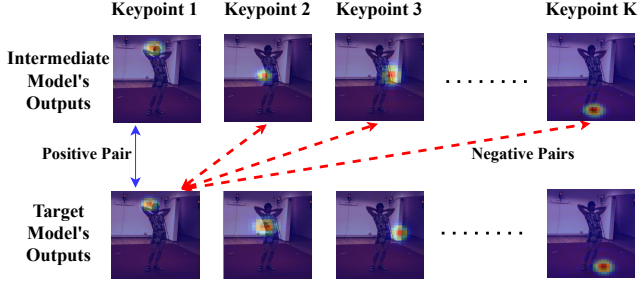
Figure 5. Illustration of building positive and negative keypoint pairs for contrastive learning between the output heatmaps of the intermediate model and target model.

In Fig. 4, we introduce the projection of a heatmap from both horizontal and vertical directions. For the heatmap $h_{i,j}^{in}$ generated by the intermediate model from ith sample's jth keypoint, its projection can be represented as $proj(h_{i,j}^{in}) = (v_{i,jx}^{in}, v_{i,jy}^{in})$, where $v_{i,jx}^{in}$ is the horizontal vector and $v_{i,jy}^{in}$ is the vertical vector. This vector pair is the foundation of the following modules and losses that we propose.

First, we propose the pose-specific contrastive loss between the intermediate model and target model. According to the observation from [12], wrong predictions are usually located at other keypoints' positions. In such a case, improving regression can be achieved by minimizing the discrepancy of heatmaps at the same location and maximizing the discrepancy of heatmaps at different locations. And that's how we define the positive and negative pair in Fig. 5. Denote the similarity of two heatmaps generated by the intermediate model and target model respectively as:

$$sim(h_{i,j}^{in}, h_{i,k}^{tg}) = \frac{1}{2}\left(\frac{v_{i,jx}^{in} \cdot v_{i,kx}^{tg}}{||v_{i,jx}^{in}||||v_{i,kx}^{tg}||} + \frac{v_{i,jy}^{in} \cdot v_{i,ky}^{tg}}{||v_{i,jy}^{in}||||v_{i,ky}^{tg}||}\right). \quad (6)$$

On the basis of the similarity, we define the overall pose-specific contrastive loss as:

$$\mathcal{L}_{cst} = -\mathbb{E}_{x_i^t \in \mathcal{T}} \mathbb{E}_{j \in \mathcal{P}} \log \frac{\exp(sim(h_{i,j}^{in}, h_{i,j}^{tg}))}{\Sigma_{k \in \mathcal{P}} \exp(sim(h_{i,j}^{in}, h_{i,k}^{tg}))}. \quad (7)$$

Next, we propose a pose-specific self-supervised technique called pose-specific information maximization. Information maximization is very helpful to source-free domain adaptation for classification [21] because it makes the target outputs individually certain and globally diverse, hence mitigating the domain gap implicitly. However, the improvements are limited due to the sparsity of heatmaps. With the help of projected vectors, we refine the information maximization loss as:

$$\mathcal{L}_{im} = \mathbb{E}_{x_i^t \in \mathcal{T}}(\mathcal{L}_{entx}(x_i^t) + \mathcal{L}_{enty}(x_i^t) - \mathcal{L}_{div}(x_i^t)), \quad (8)$$

$$\begin{cases} \mathcal{L}_{entx}(x_i^t) = -\mathbb{E}_{j \in \mathcal{P}}[\sigma(v_{i,jx}^{tg}) \log \sigma(v_{i,jx}^{tg})], \\ \mathcal{L}_{enty}(x_i^t) = -\mathbb{E}_{j \in \mathcal{P}}[\sigma(v_{i,jy}^{tg}) \log \sigma(v_{i,jy}^{tg})], \\ \mathcal{L}_{div}(x_i^t) = -[\mathbb{E}_{j \in \mathcal{P}}\sigma(h_{i,j}^{tg})] \log[\mathbb{E}_{j \in \mathcal{P}}\sigma(h_{i,j}^{tg})]. \end{cases} \quad (9)$$

Here $\mathcal{L}_{entx}$ is the entropy of horizontal vectors, and $\mathcal{L}_{enty}$ is the entropy of vertical vectors. The last term in Equation 8 is the diversity-promoting objective. Moreover, we introduce the consistency loss between the two models' outputs:

$$\mathcal{L}_{con} = \mathbb{E}_{(x_i^s, y_i^s) \in \mathcal{S}} \mathcal{L}_{mse}(F^{in}(G^{in}(x_i^s)), F^{tg}(G^{tg}(x_i^s))). \quad (10)$$

By summarizing above, we conclude the overall objective loss function for the target model as:

$$\mathcal{L}_{tgt} = \mathcal{L}_{con} + \beta\mathcal{L}_{cst} + \gamma\mathcal{L}_{im}, \quad (11)$$

where β and γ are trade-off parameters.

Moreover, the intermediate model is updated via the exponential moving average strategy (EMA) in this step:

$$\theta_{f^{in}}^{t+1} = \eta\theta_{f^{in}}^t + (1 - \eta)\theta_{f^{tg}}^t, \quad (12)$$

where t denotes the step of training and η denotes the smoothing coefficient set to 0.999 by default.

4. Experiments

Datasets. We use three human pose datasets and three hand pose datasets to validate our approach. **SURREAL** [36] is utilized as the source dataset and encompasses six million synthetic human pose images. **Human3.6M** [11], one of the target human pose datasets, is a frequently used real-world dataset in the community, consisting of 3.6 million images distributed into seven folds. , According to previous research [12, 16], S1, S5, S6, S7, and S8 are designated as the training set, whereas S9 and S11 serve as the testing set. The other target dataset is **Leeds Sports Pose** [15] (LSP), which is a real-world dataset with 2,000 images, and we use all of them for adaptation. We focus on two domain adaptation tasks: SURREAL $\rightarrow$ Human3.6M and SURREAL $\rightarrow$ LSP for human pose tasks.

Rendered Hand Pose Dataset [42] (RHD) is the source hand dataset with 43,986 synthetic hand images, of which 41,258 are allocated for training, while the remaining 2,728 images are utilized for validation. **Hand-3D-Studio** [41] (H3D) is one of the two target hand datasets, including 22,000 real-world frames. For training, we employed 18,800 frames and reserved the rest for testing, following the protocol outlined in previous research papers [12, 16]. The other target hand dataset, FreiHand [43], provides 130k real-world images, and we employed all of them for our adaptation task. Our focus is on two domain adaptation tasks: RHD $\rightarrow$ H3D and RHD $\rightarrow$ FreiHand.

Implementation Details. We adopt Simple Baseline [39] with ResNet-101 [9] as the HPE backbone, following previous works [12, 16]. In the *pretrain* process, we conduct 40 epochs, each with 500 iterations, using the Adam [17] optimizer with an initial learning rate of 1e-4. The learning rate decreases to 1e-5 at 25 epochs. For the *adaptation* process, we execute 40 epochs with 30,000 iterations. In step

A, we choose an initial learning rate of 1e-4 for F^{sr} and 1e-5 for G^{in}. Step B sets the initial learning rate to 1e-4 for G^{tg} and 1e-3 for F^{tg}. We use the same annealing strategy for the learning rate scheduler as in [12]. As for hyperparameters, we select $\alpha = 0.7$, $\beta = 0.5$, and $\gamma = 0.85$. Moreover, the intermediate model is used for the final inference.

4.1. Main Results

Baselines. Four general domain adaptive pose estimation methods **CC-SSL** [25], **MDAM** [19], **RegDA** [12] and **UniFrame** [16] with SOTA performances are chosen as baselines for comparison. We list these baselines that need to access source data to demonstrate that our proposed source-free method's performance remains competitive when compared to general domain adaptive pose estimation methods. We build source-free baselines by either adapting domain-adaptive HPE methods or source-free DA methods for classification methods. We first introduce two domain adaptive HPE methods. **RegDA-SF** is based on [12], which originally contains three steps and the first step uses source data. We replace the first step with an exponential moving averaging strategy to obtain the new method. **Uniframe-SF** is on the foundation of [16]. What we do is remove the style transfer modules between source and target so that it becomes a source-free method. The rest are adapted from source-free DA methods for classification methods. **SHOT** [21] is a classic source-free domain adaptation method for image classification, and we adapt its mutual information maximization module for HPE by treating the normalized heatmaps as output probabilities in classification. **MMT** is based on [6], which is a source-free adaptation method that applies two models mutually learning from each other. Here we replace its soft classification loss with the soft consistency loss. **SHOT++** [22] is the advanced version of [21], which uses a special unsupervised representation learning technique predicting image rotations. We also add this module to our adopted baseline.

Metrics. To evaluate the accuracy of the proposed approach, we adopt the Percentage of Correct Keypoint (PCK) metric, as used in previous work. We set the ratio of correct predictions to be 5%, and report PCK@0.05 in Tables 1-4. In addition to overall keypoint accuracy, we divide joints into several part segments and measure their performance using specific metrics. For the 21-keypoint hand skeleton, we use the Metacarpophalangeal (MCP), Proximal Interphalangeal (PIP), Distal Interphalangeal (DIP), and Fingertip (Fin) metrics. For the 18-keypoint human skeleton, we select the Shoulder (Sld), Elbow (Elb), Wrist, Hip, Knee, and Ankle metrics. These segment-specific metrics enable a more detailed evaluation of the models' performances.

Quantitative Results. We present the results of our proposed method on four different tasks: hand tasks RHD→H3D and RHD→FreiHand, and human tasks

Table 1. PCK@0.05 on RHD → H3D Task

Method	SF	MCP	PIP	DIP	Fin	All
Source-only	-	67.4	64.2	63.3	54.8	61.8
CC-SSL [25] (CVPR'20)	×	81.5	79.9	74.4	64.0	75.1
MDAM [19] (CVPR'21)	×	82.3	79.6	72.3	61.5	74.1
RegDA [12] (CVPR'21)	×	79.6	74.4	71.2	62.9	72.5
UniFrame [16] (ECCV'22)	×	86.7	84.6	78.9	68.1	79.6
RegDA-SF [12] (CVPR'21)	✓	71.2	66.8	66.2	58.5	66.9
UniFrame-SF [16] (ECCV'22)	✓	84.8	84.2	77.0	68.0	77.9
SHOT [21] (ICML'20)	✓	77.5	69.8	70.6	64.2	73.7
MMT [6] (ICLR'20)	✓	78.7	74.1	72.5	65.3	75.3
SHOT++ [22] (TPAMI'22)	✓	85.3	85.1	78.2	67.6	78.9
Ours	✓	**88.4**	**89.2**	**80.9**	**71.4**	**82.2**

Table 2. PCK@0.05 on RHD → FreiHand Task

Method	SF	MCP	PIP	DIP	Fin	All
Source-only	-	35.2	50.1	54.8	50.7	46.8
CC-SSL [25] (CVPR'20)	×	37.4	48.2	50.1	46.5	43.8
MDAM [19] (CVPR'21)	×	32.3	48.1	51.7	47.3	45.1
RegDA [12] (CVPR'21)	×	40.9	55.0	58.2	53.1	51.1
UniFrame [16] (ECCV'22)	×	43.5	64.0	67.4	62.4	58.5
RegDA-SF [12] (CVPR'21)	✓	38.6	52.9	57.6	54.3	49.5
UniFrame-SF [16] (ECCV'22)	✓	40.6	62.5	61.0	60.2	55.7
SHOT [21] (ICML'20)	✓	40.4	61.3	60.5	58.1	53.3
MMT [6] (ICLR'20)	✓	39.6	60.4	60.0	57.8	52.6
SHOT++ [22] (TPAMI'22)	✓	41.0	62.8	62.7	59.3	55.8
Ours	✓	**43.7**	**65.9**	**66.6**	**63.1**	**58.8**

Table 3. PCK@0.05 on SURREAL → Human3.6M Task

Method	SF	Sld	Elb	Wrist	Hip	Knee	Ankle	All
Source-only	-	69.4	75.4	66.4	37.9	77.3	77.7	67.3
CC-SSL [25] (CVPR'20)	×	44.3	68.5	55.2	22.2	62.3	57.8	51.7
MDAM [19] (CVPR'21)	×	51.7	83.1	68.9	17.7	79.4	76.6	62.9
RegDA [12] (CVPR'21)	×	73.3	86.4	72.8	54.8	82.0	84.4	75.6
UniFrame [16] (ECCV'22)	×	78.1	89.6	81.1	52.6	85.3	87.1	79.0
RegDA-SF [12] (CVPR'21)	✓	70.6	82.0	69.8	43.3	79.1	79.4	71.5
UniFrame-SF [16] (ECCV'22)	✓	74.3	84.8	72.5	45.9	81.5	85.2	75.1
SHOT [21] (ICML'20)	✓	73.0	83.2	71.9	46.7	79.8	82.4	73.3
MMT [6] (ICLR'20)	✓	73.2	83.5	72.4	45.1	80.8	83.9	73.9
SHOT++ [22] (TPAMI'22)	✓	75.0	85.2	76.7	45.3	82.8	85.0	75.3
Ours	✓	**77.9**	**88.8**	**80.4**	**52.3**	**84.2**	**86.9**	**78.7**

Table 4. PCK@0.05 on SURREAL → LSP Task

Method	SF	Sld	Elb	Wrist	Hip	Knee	Ankle	All
Source-only	-	51.5	65.0	62.9	68.0	68.7	67.4	63.9
CC-SSL [25] (CVPR'20)	×	36.8	66.3	63.9	59.6	67.3	70.4	60.7
MDAM [19] (CVPR'21)	×	61.4	77.7	75.5	65.8	76.7	78.3	69.2
RegDA [12] (CVPR'21)	×	62.7	76.7	71.1	81.0	80.3	75.3	74.6
UniFrame [16] (ECCV'22)	×	69.2	84.9	83.3	85.5	84.7	84.3	82.0
RegDA-SF [12] (CVPR'21)	✓	54.8	70.5	67.6	65.4	73.2	70.0	66.5
UniFrame-SF [16] (ECCV'22)	✓	68.4	80.5	79.1	82.7	80.8	81.0	78.8
SHOT [21] (ICML'20)	✓	63.5	72.7	66.5	78.4	79.7	73.2	72.4
MMT [6] (ICLR'20)	✓	60.9	70.9	70.3	81.1	79.3	72.8	71.5
SHOT++ [22] (TPAMI'22)	✓	69.5	81.7	80.9	84.0	82.3	79.7	79.9
Ours	✓	**70.7**	**85.4**	**83.8**	**86.6**	**85.2**	**85.0**	**83.2**

SURREAL→Human3.6M and SURREAL→LSP. Tables 1 and 2 show the results for the hand tasks, where our method achieves state-of-the-art performance, surpassing the second best baseline SHOT++ with a significant lead of 3.3% and 3.0%, respectively. Moreover, our method outperforms the SHOT++ model by 4.1% and 3.9% on the PIP metric, respectively. For the human tasks, Tables 3 and 4 display the results. Our model achieves a considerable margin of 3.4% and 3.3% over the second best baseline SHOT++ for SURREAL→Human3.6M and SURREAL→LSP, respectively. Specifically, our method outperforms SHOT++ by 7.0% on the Wrist for the adaptation on Human3.6M and 5.3% on the Ankle for the adaptation on LSP. Furthermore, our approach remains competitive when compared with general domain adaptive HPE methods that need to access source data, as it surpasses UniFrame by 2.6%
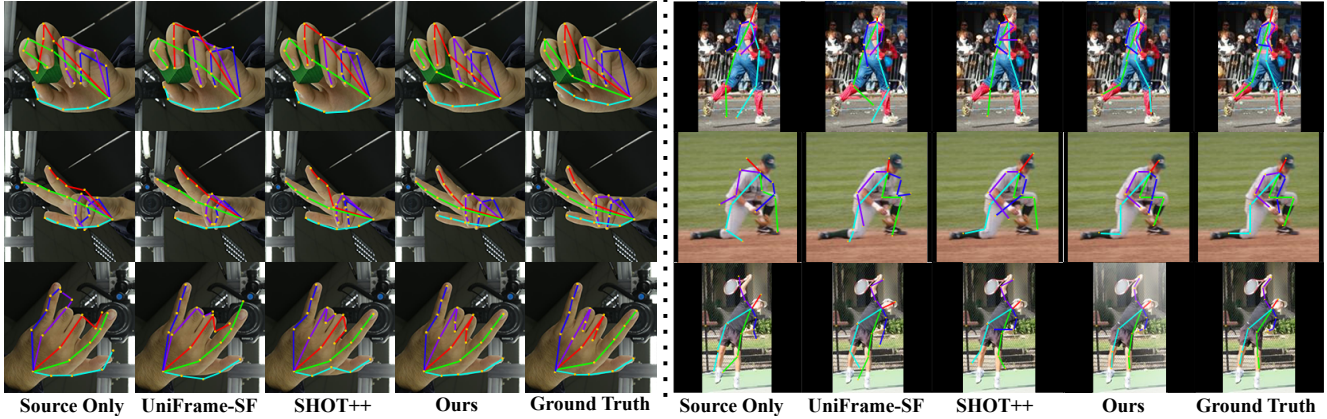
| Source Only | UniFrame-SF | SHOT++ | Ours | Ground Truth | Source Only | UniFrame-SF | SHOT++ | Ours | Ground Truth |

Figure 6. Qualitative results on H3D dataset (left) and LSP dataset (right).

on RHD→H3D, 0.3% on RHD→FreiHand, and 1.2% on SURREAL→LSP.

Qualitative Results. Figs. 6 shows qualitative results of H3D on the left side and LSP on the right side. We use Source only, **UniFrame-SF** [16], **SHOT++** [22], **Ours**, and **Ground Truth** for qualitative comparison. It is evident that our method outperforms other baselines significantly.

4.2. Ablation Study on Framework

Our method contains two modules in Step A & Step B as the source-protect module (**SP**) and the target-relevant module (**TR**) separately, and here we focus on their functions. Moreover, we use **MMT** [6] as the baseline, which is referred in Table 1-4 . Table 5 and 6 show the ablation study of frameworks on RHD→H3D and SURREAL→Human3.6M with different combinations of these modules.

Table 5. Ablation of Frameworks on RHD → H3D

Method	MCP	PIP	DIP	Fin	All
MMT [6]	78.7	74.1	72.5	65.3	75.3
MMT [6] +TR	81.3	77.5	76.0	67.7	79.1
SP+TR (Ours)	88.4	89.2	80.9	71.4	82.2

Table 6. Ablation of Frameworks on SURREAL → Human3.6M

Method	Sld	Elb	Wrist	Hip	Knee	Ankle	All
MMT [6]	73.2	83.5	72.4	45.1	80.8	83.9	73.9
MMT [6] +TR	74.4	86.5	75.7	48.9	81.3	85.1	76.2
SP+TR (Ours)	77.9	88.8	80.4	52.3	84.2	86.9	78.7

The results clearly demonstrate that both SP and TR contribute to improving the model's performance. Specifically, TR enhances the model's accuracy by 3.8% on RHD → H3D and 2.3% on SURREAL → Human3.6M, while SP leads to an improvement of 4.1% on RHD → H3D and 2.5% on SURREAL → Human3.6M compared to MMT. Notably, the two proposed modules provide similar levels of improvement.

4.3. Ablation Study on Proposed Losses

We performed a detailed ablation study on the three proposed losses, namely $\mathcal{L}_{res}$, $\mathcal{L}_{cst}$, and $\mathcal{L}_{im}$, using the RHD→H3D and SURREAL→Human3.6M tasks. Tables 7 and 8 present the results. As we have previously discussed the functions and advantages of the source-protect and target-relevant modules in Sec. 4.2, we retained $\mathcal{L}_{ft}$ from the source-protect module and $\mathcal{L}_{con}$ from the target-relevant module in the **Baseline**.

Table 7. Ablation of Losses on RHD → H3D

Method	MCP	PIP	DIP	Fin	All
Baseline	85.3	86.8	78.5	68.1	79.4
$\mathcal{L}_{res}$	85.5	87.9	79.3	70.7	80.1
$\mathcal{L}_{cst}$	85.5	89.0	78.4	70.1	80.9
$\mathcal{L}_{im}$	85.8	87.2	78.7	69.8	80.3
$\mathcal{L}_{cst}\&\mathcal{L}_{im}$	86.9	88.8	80.1	70.3	81.4
$\mathcal{L}_{res}\&\mathcal{L}_{cst}\&\mathcal{L}_{im}$	88.4	89.2	80.9	71.4	82.2

Table 8. Ablation of Losses on SURREAL → Human3.6M

Method	Sld	Elb	Wrist	Hip	Knee	Ankle	All
Baseline	75.4	86.1	76.8	46.5	83.0	85.0	75.8
$\mathcal{L}_{res}$	76.0	86.9	77.5	47.8	83.3	85.2	76.6
$\mathcal{L}_{cst}$	76.6	87.4	78.8	50.6	83.5	86.0	77.2
$\mathcal{L}_{im}$	76.2	87.1	78.3	49.7	83.2	85.7	76.9
$\mathcal{L}_{cst}\&\mathcal{L}_{im}$	77.3	87.7	79.2	51.2	83.6	86.4	78.0
$\mathcal{L}_{res}\&\mathcal{L}_{cst}\&\mathcal{L}_{im}$	77.9	88.8	80.4	52.3	84.2	86.9	78.7

We observe that each loss is able to boost the model's performance. Simply applying $\mathcal{L}_{res}$ leads to a increase of 0.7% in RHD→H3D and 0.8% in SURREAL→Human3.6M. $\mathcal{L}_{cst}$ causes an improvement of 1.5% in RHD→H3D and 1.4% in SURREAL→Human3.6M. As for $\mathcal{L}_{im}$, adding it causes a rise of 0.9% in RHD→H3D and 1.1% in SURREAL→Human3.6M. Besides, it is noticed that $\mathcal{L}_{cst}$ plays a more important role than $\mathcal{L}_{res}$ or $\mathcal{L}_{im}$.

4.4. Ablation of Sparsity Reduction

In this section, we investigate the effectiveness of our proposed sparsity reduction strategy. Specifically, we compare heatmap-based methods and vector-based methods for two loss functions: contrastive learning ($\mathcal{L}_{cst}$) and information maximization ($\mathcal{L}_{im}$). We denote heatmap-based contrastive learning and information maximization as **HBCL** and **HBIM**, respectively, and vector-based contrastive learning and information maximization as **VBCL** and **VBIM**, respectively. Two tasks are selected and the results of these comparisons are presented in Tab. 9 and Tab. 10.

Table 9. Ablation of Sparsity Reduction on RHD $\to$ H3D

Method	MCP	PIP	DIP	Fin	All
MMT [6]	78.7	74.1	72.5	65.3	75.3
MMT [6] + HBCL	77.4	73.0	70.7	65.5	74.2
MMT [6] + VBCL	**81.7**	**76.6**	**74.4**	**67.9**	**78.7**
MMT [6] + HBIM	79.3	74.7	72.9	65.8	75.9
MMT [6] + VBIM	**80.3**	**76.1**	**73.6**	**66.5**	**77.5**

Table 10. Ablation of Sparsity Reduction on SURREAL $\to$ Human3.6M

Method	Sld	Elb	Wrist	Hip	Knee	Ankle	All
MMT [6]	73.2	83.5	72.4	45.1	80.8	83.9	73.9
MMT [6] + HBCL	71.7	82.0	71.8	44.6	80.2	82.4	73.0
MMT [6] + VBCL	**75.6**	**87.0**	**75.1**	**49.2**	**83.3**	**86.3**	**76.7**
MMT [6] + HBIM	73.9	84.6	72.7	46.0	81.1	83.0	74.3
MMT [6] + VBIM	**75.3**	**86.4**	**74.3**	**48.6**	**82.4**	**85.7**	**76.0**

Based on the results, it is evident that HBCL does not improve the model's performance significantly, whereas VBCL provides significant assistance. For instance, MMT+HBCL shows a decrease of 1.1% on the RHD $\to$ H3D task, while MMT+VBCL exhibits an increase of 3.4%. Additionally, while HBIM can enhance the model's performance, it falls short of surpassing VBIM. For example, MMT+HBIM shows an increase of 0.6% on the RHD $\to$ H3D task, whereas MMT+VBIM achieves an increase of 2.2%. In summary, it becomes evident that utilizing the vector-based spatial probability space is crucial for achieving optimal results.
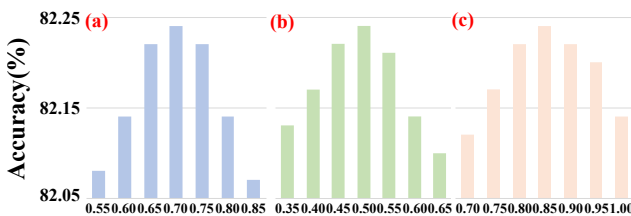
4.5. Parameter Analysis

Figure 7. Parameter Analysis on RHD$\to$H3D (best viewed in color). **a**: Analysis on α. **b**: Analysis on β. **c**: Analysis on γ.

We use RHD$\to$H3D to illustrate the sensitivity of α in Equation 5 and β and γ in Equation 11. The results are shown in Fig. 7. From it, we observe that $\alpha = 0.7$, $\beta = 0.5$, and $\gamma = 0.85$ is the best choice. Moreover, it is noticed that the decreases from different parameters are limited to 0.2% for α, β and γ. Therefore, our method shows stable performance over hyperparameters.

5. Qualitative Results on the Unseen Dataset COCO

Figure 8. Qualitative results on the COCO dataset. Here we compare our method trained from SURREAL $\to$ LSP with the source-only model.

In this section, we showcase qualitative results on the COCO dataset [23], which was not utilized in our previous human dataset studies. It's important to note that our 14-keypoint poses are annotated differently from the 17-keypoint poses in COCO, which hinders providing quantitative results on COCO using its own metrics or APIs. Despite this limitation, we believe that our qualitative findings hold significant value. For our analysis, we leverage the model trained on the SURREAL $\to$ LSP task in conjunction with the pre-trained SURREAL model. The results, presented in Fig. 8, demonstrate the model's capacity to generalize effectively to unseen datasets, implying that the adaptation process boosts the model's generalization ability beyond its original training dataset.

6. Conclusion

In this paper, we propose a new task named source-free domain adaptive human pose estimation that places an emphasis on the privacy of source data. Additionally, we propose a new framework that includes source-protect and target-relevant modules, which aim to alleviate the issues of catastrophic forgetting of source and the sparsity of spatial distributions, respectively. Our approach is evaluated on hand and human pose datasets through extensive experiments, demonstrating that it outperforms state-of-the-art methods by a considerable margin.

References

[1] Dian Chen, Dequan Wang, Trevor Darrell, and Sayna Ebrahimi. Contrastive test-time adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 295–305, June 2022.

[2] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020.

[3] Yilun Chen, Zhicheng Wang, Yuxiang Peng, Zhiqiang Zhang, Gang Yu, and Jian Sun. Cascaded pyramid network for multi-person pose estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7103–7112, 2018.

[4] Wanxia Deng, Qing Liao, Lingjun Zhao, Deke Guo, Gangyao Kuang, Dewen Hu, and Li Liu. Joint clustering and discriminative feature alignment for unsupervised domain adaptation. *IEEE Transactions on Image Processing*, pages 7842–7855, 2021.

[5] Yaroslav Ganin and Victor Lempitsky. Unsupervised domain adaptation by backpropagation. In *International Conference on Machine Learning*, pages 1180–1189, 2015.

[6] Yixiao Ge, Dapeng Chen, and Hongsheng Li. Mutual mean-teaching: Pseudo label refinery for unsupervised domain adaptation on person re-identification. In *International Conference on Learning Representations*, 2020.

[7] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020.

[8] Zhongyi Han, Haoliang Sun, and Yilong Yin. Learning transferable parameters for unsupervised domain adaptation. *IEEE Transactions on Image Processing*, 31:6424–6439, 2022.

[9] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016.

[10] Xun Huang and Serge Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1501–1510, 2017.

[11] Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. Human3. 6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(7):1325–1339, 2013.

[12] Junguang Jiang, Yifei Ji, Ximei Wang, Yufeng Liu, Jianmin Wang, and Mingsheng Long. Regressive domain adaptation for unsupervised keypoint detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6780–6789, 2021.

[13] Ruijie Jiang, Thuan Nguyen, Prakash Ishwar, and Shuchin Aeron. Supervised contrastive learning with hard negative samples. *arXiv preprint arXiv:2209.00078*, 2022.

[14] Rui Jin, Jing Zhang, Jianyu Yang, and Dacheng Tao. Multi-branch adversarial regression for domain adaptive hand pose estimation. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(9):6125–6136, 2022.

[15] Sam Johnson and Mark Everingham. Clustered pose and nonlinear appearance models for human pose estimation. In *British Machine Vision Conference*, volume 2, page 5. Citeseer, 2010.

[16] Donghyun Kim, Kaihong Wang, Kate Saenko, Margrit Betke, and Stan Sclaroff. A unified framework for domain adaptive pose estimation. *arXiv preprint arXiv:2204.00172*, 2022.

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

[18] Vinod K Kurmi, Venkatesh K Subramanian, Vinay P Namboodiri, , and . Domain impression: A source data free domain adaptation method. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 615–625, 2021.

[19] Chen Li and Gim Hee Lee. From synthetic to real: Unsupervised domain adaptation for animal pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1482–1491, 2021.

[20] Rui Li, Qianfen Jiao, Wenming Cao, Hau-San Wong, and Si Wu. Model adaptation: Unsupervised domain adaptation without source data. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9638–9647, 2020.

[21] Jian Liang, Dapeng Hu, and Jiashi Feng. Do we really need to access the source data? source hypothesis transfer for unsupervised domain adaptation. In *International Conference on Machine Learning*, pages 6028–6039. PMLR, 2020.

[22] Jian Liang, Dapeng Hu, Yunbo Wang, Ran He, and Jiashi Feng. Source data-absent unsupervised domain adaptation through hypothesis transfer and labeling transfer. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(11):8602–8617, 2022.

[23] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European Conference on Computer Vision*, pages 740–755. Springer, 2014.

[24] Mingsheng Long, Yue Cao, Jianmin Wang, and Michael Jordan. Learning transferable features with deep adaptation networks. In *International Conference on Machine Learning*, pages 97–105, 2015.

[25] 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*, pages 12386–12395, 2020.

[26] Qucheng Peng, Zhengming Ding, Lingjuan Lyu, Lichao Sun, and Chen Chen. Toward better target representation for source-free and black-box domain adaptation. *arXiv preprint arXiv:2208.10531*, 2022.

[27] Ekkasit Pinyoanuntapong, Ayman Ali, Kalvik Jakkala, Pu Wang, Minwoo Lee, Qucheng Peng, Chen Chen, and Zhi Sun. Gaitsada: Self-aligned domain adaptation for mmwave gait recognition. *arXiv preprint arXiv:2301.13384*, 2023.

[28] 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*, pages 3723–3732, 2018.

[29] Dahu Shi, Xing Wei, Liangqi Li, Ye Ren, and Wenming Tan. End-to-end multi-person pose estimation with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11069–11078, 2022.

[30] Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang. Deep high-resolution representation learning for human pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5693–5703, 2019.

[31] Hui Tang, Ke Chen, and Kui Jia. Unsupervised domain adaptation via structurally regularized deep clustering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8725–8735, 2020.

[32] Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. *Advances in Neural Information Processing Systems*, 30, 2017.

[33] Jonathan Tompson, Ross Goroshin, Arjun Jain, Yann LeCun, and Christoph Bregler. Efficient object localization using convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 648–656, 2015.

[34] Jonathan J Tompson, Arjun Jain, Yann LeCun, and Christoph Bregler. Joint training of a convolutional network and a graphical model for human pose estimation. *Advances in neural information processing systems*, 27, 2014.

[35] Eric Tzeng, Judy Hoffman, Ning Zhang, Kate Saenko, and Trevor Darrell. Deep domain confusion: Maximizing for domain invariance. *arXiv preprint arXiv:1412.3474*, 2014.

[36] Gul Varol, Javier Romero, Xavier Martin, Naureen Mahmood, Michael J Black, Ivan Laptev, and Cordelia Schmid. Learning from synthetic humans. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 109–117, 2017.

[37] Haixin Wang, Lu Zhou, Yingying Chen, Ming Tang, and Jinqiao Wang. Regularizing vector embedding in bottom-up human pose estimation. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part VI*, pages 107–122. Springer, 2022.

[38] Haifeng Xia, Handong Zhao, and Zhengming Ding. Adaptive adversarial network for source-free domain adaptation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9010–9019, 2021.

[39] Bin Xiao, Haiping Wu, and Yichen Wei. Simple baselines for human pose estimation and tracking. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 466–481, 2018.

[40] Feng Zhang, Xiatian Zhu, Hanbin Dai, Mao Ye, and Ce 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.

[41] Zhengyi Zhao, Tianyao Wang, Siyu Xia, and Yangang Wang. Hand-3d-studio: A new multi-view system for 3d hand reconstruction. In *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2478–2482, 2020.

[42] Christian Zimmermann and Thomas Brox. Learning to estimate 3d hand pose from single rgb images. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4903–4911, 2017.

[43] Christian Zimmermann, Duygu Ceylan, Jimei Yang, Bryan Russell, Max Argus, and Thomas Brox. Freihand: A dataset for markerless capture of hand pose and shape from single rgb images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 813–822, 2019.