ACTUPose: Active Curriculum Training for Unsupervised Domain Adaptation in Pose Estimation

Anonymous CVPR submission

Paper ID *****

Abstract

In this study, we present a novel training methodology for unsupervised domain adaptation (UDA) in the context of pose estimation. Existing UDA methods for pose estimation often struggle to generalize effectively across similar poses in the target data, even when such poses are well-represented in the source data. We attribute this challenge to the lack of uniform support for different poses and a systematic training strategy for handling poses of varying complexity in the source domain. To address this challenge, we propose ACTUPose, an active curriculum training strategy that utilizes the diversity of poses in the source data. Within this framework we incorporate a method for quantifying pose complexity that dynamically orders the training data as the training progresses. We further introduce a new loss function aimed at improving skeletal structure prediction. Additionally, we incorporate a cross-domain feature loss to better utilize unlabeled real data. With this approach we demonstrate state-of-the-art performance across standard benchmark datasets for UDA in Pose Estimation.

1. Introduction

Pose estimation is an important problem in computer vision with applications ranging from autonomous driving [35, 36], motion capture [8], and robotics [38]. However, despite remarkable progress, these tasks remain challenging, primarily due to the inherent variability in poses across different domains, environments, and data sources. The robustness and accuracy of pose estimation models heavily rely on the availability of large and diverse annotated datasets for training. To this end, the synthesis of labeled training data using modern computer graphics[24, 30] is becoming increasingly relevant, allowing the creation of vast datasets under controlled conditions. These synthetic datasets reduce the need for laborious manual keypoint annotations. However, pose estimation models trained on rendered synthetic data suffer from a domain gap problem,

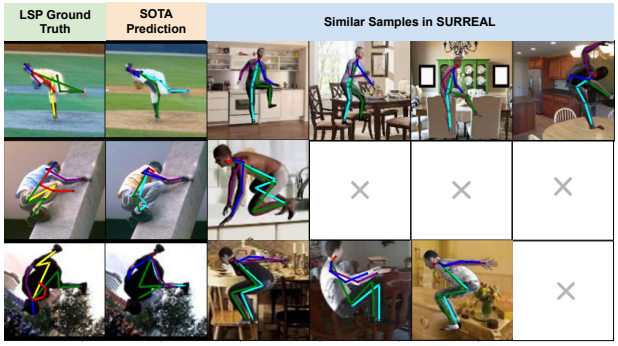

Figure 1. LSP [15] ground truth, prediction by state-of-the-art (SOTA): UDAPE [17] and similar poses in SURREAL [30]. Despite SURREAL's diverse scenarios, UDAPE faces challenges in effective generalization. Could this be linked to the uneven representation of poses in the source domain?

arising from the differences in appearance, viewpoint and lighting conditions. This can significantly affect the performance of model to new, unseen domains. A specific gap more relevant to pose estimation is the 'pose distribution gap', a gap between available poses in the source and target domains. Pose distribution gap under unsupervised domain adaptation (UDA) has been scarcely studied in the literature, which is being addressed in this paper. With a careful analysis of the state-of-the-art UDA models for pose estimation, e.g., UDAPE [17], we noticed a significant limitation that they struggle to generalize across poses in the target domain although similar poses were present in the source domain. As shown in Figure 1, there are similar poses present in source domain SURREAL [30], albeit with difference in 3D orientation. However, the UDAPE model fails to predict these poses accurately. We hypothesize, that this failure to generalize originates from the lack of uniform support across poses of varying complexity in the source domain. We support this hypothesis through a careful analysis of the source domain poses as shown in Figure 3. Here, we sort and categorize the source domain poses based on a com-

plexity measure and observe that as we move towards the more difficult categories, the variability of poses increases indicating lower support. These observations motivated us to investigate methods to (a) categorize the source pose distribution, (b) strategically use this categorization for training and, (c) improve the generalization of pose estimation methods trained on synthetic data towards the target domain for the first time.

Existing efforts to categorize the pose distribution [12], [5] fail to accurately represent the skeletal geometry as they account for each joint independently but do not model the plausibility of the overall pose. In contrast to the above methods, we introduce a method, to score the source domain poses using an auxiliary deep learning model and categorize the poses based on this score. We train the auxiliary scoring model as a variational auto-encoder whose input and output supervision is the same 2D skeleton. The inability of the model to generalize towards some poses during testing helps estimate the complexity of the pose in the source distribution. This helps us to score and categorize the source pose distribution into closely related groups.

The above mentioned categorization presents the challenge of selecting samples effectively for training. To overcome this challenge, we propose a novel active curriculum learning strategy. The curriculum enables the model to choose the samples dynamically based on the model training status. Finally, we primarily want to improve the performance of the model on the target domain data. We therefore use an existing unsupervised domain adaptation technique [17] and introduce additional guidance to the model in the form of new loss functions to generalize better to the target domain. The first loss function that we introduce, aids in preserving the geometry of the pose. The second loss function that we introduce, aims to minimize the feature gap between similar poses in the source and target domains.

We conduct extensive experiments on UDA benchmarks and summarize our key contributions as follows:

- **Pose Analysis:** We use a VAE-based scoring mechanism[9] to assess pose complexity, enabling efficient sorting and categorization.
- **Active Curriculum Learning:** We propose a novel active curriculum learning strategy to strategically use the categorized poses for training, ensuring uniform utilization of poses of varying complexity.
- **Enhanced UDA Losses:** We extend [17] with two new losses—one maintaining pose geometry and another aligning source-target features—to improve domain adaptation.
- **Performace Gain:** Our method improves accuracy by 2.3% on LSP[15], 4.5% on the Human3.6M[13] dataset, and 1.1% on H3D[39] dataset.

Together we refer to the above contributions as our method ACTUPose.

2. Related Work

Curriculum Learning is a training paradigm that structures sample presentation based on difficulty. Introduced by [2], it enhances convergence and generalization by sequencing training samples effectively. A related approach, **Self-Paced Learning** [19], prioritizes easy examples first, gradually incorporating harder ones. Unlike these methods, our strategy integrates both external difficulty scores and model performance for dynamic training adjustments. Curriculum learning has been applied across various tasks, including semi-supervised image classification [10], language modeling [11], weakly-supervised object detection [32, 37], localization [21, 29], person re-identification [31], semantic segmentation [1, 6], and image generation [27], but remains underexplored in **human pose estimation**. Notable efforts include a multi-stage curriculum strategy [25] that progressively trains on easier poses before introducing complex ones, improving 3D pose estimation in challenging cases. Another approach [7] defines difficulty using dataset statistics and multi-model evaluations. Building on these, our work introduces a more adaptive curriculum that not only categorizes poses but dynamically selects training samples based on real-time model feedback, improving generalization to complex and unseen poses. While curriculum learning has been explored in source-free domain adaptation [4, 16], its application in source-to-target UDA remains largely untapped.

Self-Training with pseudo labels Self-training with pseudo-labels involves iteratively training a model on labeled data, generating pseudo-labels for unlabeled data, and fine-tuning on the combined set to improve performance in semi-supervised or unsupervised learning. Some works which prove the efficacy of self training using pseudo labels includes [3, 22, 23]. In the context of UDA, there are a few works utilizing self-training for classification, segmentation, and detection tasks. For example, [41, 42] utilized self-training for classification and segmentation adaptation. [18] utilized weak self-training for single-shot detector adaptation. [26] used pseudo-labeled data but the labels are from extra video data annotation, which belongs to weakly-supervised domain adaptation. For pose estimation, UDAPE[17] uses self training to enforce consistency between student and teacher model, using target domain pseudo labels. On the contrary, we propose a cross domain loss using both source and target domain pseudo labels. Additionally, we improve the consistency loss used in UDAPE [17] by adding skeleton structure information derived from target pseudo labels. There are some existing efforts to categorize the human pose distribution, using joint distances or similar heuristics and k-means clustering [12]. These neither work for very large dataset such as SURREAL[30], nor do they provide a rank or order to the created clusters. To address this, we use Pose-VAE[9] to categorize and rank

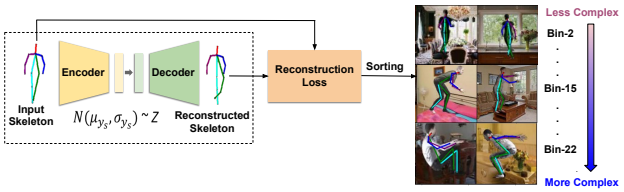

Figure 2. Pose VAE is utilized to categorize the source data into k bins in the order of increasing complexity.

samples into n bins.

3. ACTUPose

UDA is characterised by a labeled source dataset $S = \{(x_s^i, y_s^i)\}_{i=1}^N$, where x_s is the source image, y_s is the corresponding keypoint annotation and N is the number of source images and an unlabeled target dataset $T = \{(x_t^i)\}_{i=1}^M$, where x_t is the target image and M is the number of target images. We train a keypoint regression model for pose estimation using the source dataset in a supervised manner, and utilize the target domain data in an unsupervised manner to transfer the learned knowledge to the target domain. Figure 4, visualizes our proposed approach called ACTUPose. The proposed approach follows an active curriculum strategy to train the keypoint regression model. This involves first sorting the data based on a difficulty score and subsequently, training the model with a curricular strategy which actively guides and samples relevant data from the source domain (Section 3.1). We then use an unsupervised domain adaptation method to train on the unlabeled target data (Section 3.2). Both these components and the training procedure is described in detail in the following subsections.

3.1. Curriculum Training

In this section, we describe the proposed active curriculum strategy for guiding source domain training of the model. We first introduce a scoring and sorting mechanism and subsequently discuss the curriculum strategy.

3.1.1. Pose VAE

We train a variational auto-encoder (VAE) to estimate the complexity of a pose in the source dataset, to rank the data. The input and output supervision for the VAE is the same set of 2D keypoints. These keypoints are flattened out and passed as input to the VAE. The VAE follows an encoder–decoder architecture, as illustrated in Figure 2. The encoder produces the latent space parameters: a mean and log-variance vector. The decoder then samples from a normal distribution parameterized by these two values. The latent distribution is regularized using the Kullback–Leibler divergence loss. The loss function used to train the VAE is mentioned below in Equation 1.

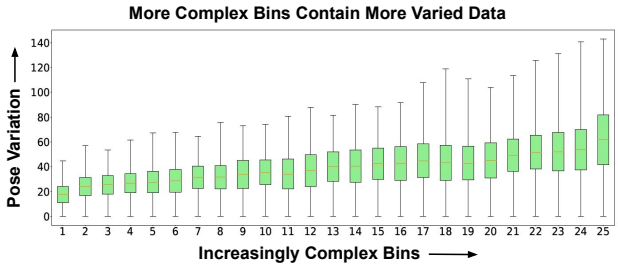

Figure 3. The box plot shows the mean angular difference between poses within each bin, representing pose variation. As the bins progress towards more complex poses, both the mean and variance increase. A higher mean indicates greater deviation from supported poses, while a higher variance reflects increased diversity of poses within the bin.

$$\mathcal{L}_{\text{VAE}} = \sum_{i=0}^{2K} \|y_s^i - \hat{y}_s^i\|^2 + \lambda \, \text{KL}\left[\mathcal{N}(\mu_s, \sigma_s) \,\|\, \mathcal{N}(0, I)\right]. \quad (1)$$

Here, y_s^i denotes the ith keypoint location for a pose sampled from the source domain, K is the number of keypoints in a pose, $\hat{y}_s^i$ is the prediction of the VAE decoder for the ith keypoint, KL denotes the Kulback-Leibler divergence loss, $\mathcal{N}$ is a normal distribution with μ_s and σ_s as mean and standard deviation of the VAE's latent space. The trained VAE enables us to compute the reconstruction error as a mean-squared error as defined in Equation 2, for a given pose y_s and VAE output $\hat{y}_s$:

$$\text{score} = \sum_{i=0}^{2K} \|y_s^i - \hat{y}_s^i\|^2. \quad (2)$$

If the reconstruction error is high, the model is unable to reconstruct a pose. This provides a useful mechanism to score poses in the source domain. Using this score, we categorize the source poses into a fixed number of bins. To achieve this, we sort all poses from the least to the largest reconstruction error and then uniformly divide them into N_b bins, each containing same number of poses. We hypothesize that the reconstruction error is a measure of pose complexity and limited support for such poses in the dataset. This hypothesis is supported in Figure 3. Here, we compute the variability of poses within a bin using the mean angular difference. The measure is defined in equation 6 of Section 3.4, where we define to identify suitable pairs to apply a cross-domain loss, here we apply it to illustrate bin characteristics.

3.1.2. Active Curriculum

As discussed in the previous subsection, the data is categorized into bins based on the pose complexity. In order to use this data to train a pose estimation model, we progressively train from the low complexity to the high complexity poses. The amount of data sampled from each bin depends

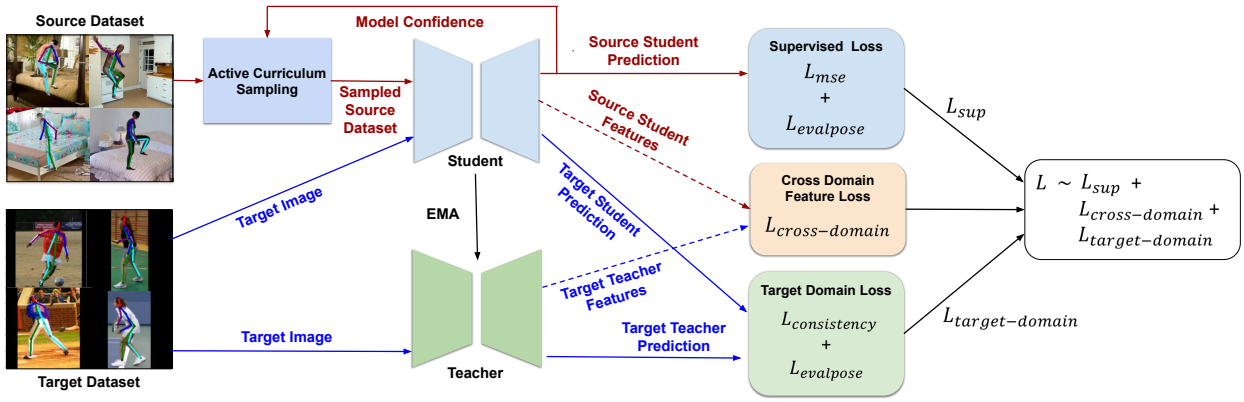

Figure 4. ACTUPose: The source data is chosen using the Active Curriculum Sampling for each epoch. The student model is trained using this sampled data in a supervised manner. The target dataset is utilised to train the student model using target domain loss applied between student and teacher predictions along with cross domain feature loss applied between the features of similar source and target poses. Note that L_{mse} and $L_{consistency}$ are same as used in baseline model, UDAPE[17].

on the current training epoch. Their relative contribution towards training the model, shifts towards the more complex bins as the training progresses. The sampling follows a soft training schedule based on a Gaussian distribution centered around a chosen bin. This is in contrast to a conventional curriculum schedule that chooses samples from a single bin for an entire training epoch. This distribution schedule is visualized in Figure 5. The probability p_j of sampling from a bin j is defined in the Equation 3,

$$p_j = \exp\left(\frac{-(s_j - \mu_g)^2}{2(\sigma s_j)^2}\right), \quad j \in \{0 \dots N_b\}. \tag{3}$$

Here, N_b is the total number of bins, μ_g is the mean score for bin g, around which the Gaussian has its peak for the epoch e. g refers to the bin with the highest contribution in epoch e, $g = min(\lambda N_b e/epochs, N_b)$. We refer to λ as a pacing parameter that controls the rate at which the Gaussian curve's peak moves towards the harder bins. $epochs$ refers to the total number of epochs the model is trained for. The score s_j in equation 3 is the mean reconstruction error for a bin j. Total number of points sampled from a bin is therefore $p_j(N/N_b)$.

To make this sampling schedule sensitive to the current ability of the pose estimation model to predict a pose, we incorporate the model's confidence into the score computation. The motivation is to re-rank the source domain images in accordance to how difficult it is for the model to predict the pose in the image. The updated score for a sample is computed using the Equation 4:

$$\text{score} = \frac{K}{\sum_{i=0}^{K} \max(\hat{H}_s^i)} \cdot \text{score}. \tag{4}$$

where, $\hat{H}_s^i$ is the predicted heatmap [1] for the ith keypoint.

[1]we follow the standard pose estimation model training of predicted $K \times H \times W$ heatmaps as in [33]

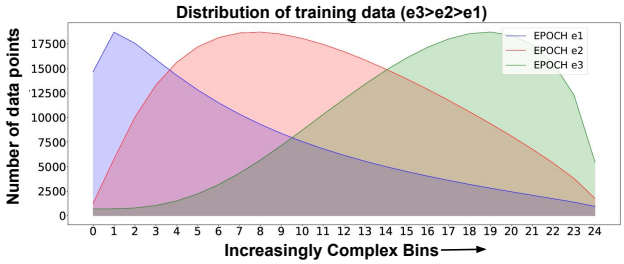

Figure 5. Active sample selection across epochs and bins: Visualizing data distribution evolution. Blue, red, and green represent the data samples distribution at different epochs. The area under the curves represent the data sampled for these epochs.

The score is essentially weighted by the average model prediction confidence, from the previous epoch of training. Intuitively, the higher the confidence, the lower the score, and hence the better the model's understanding of the source data point. On the contrary, a lower confidence indicates that the model finds it difficult to estimate the pose for a source data point. At the end of each epoch, the source data is sorted based on the updated scores and reassigned to bins, following the schedule in Equation 3. This reranking process is illustrated in Figure 6.

3.2. Unsupervised Domain Adaptation

The previous subsection introduced our active curriculum training strategy for training the model with the source domain data. However, we also use the unlabeled target dataset T, within an unsupervised domain adaptation (UDA) setting. For domain adaptation, we use an existing state-of-the-art approach UDAPE [17]. UDAPE follows a student-teacher learning paradigm for pose estimation. It additionally uses a consistency loss on pseudo-labels and bridges the appearance level gap using a bi-directional style transfer method. We introduce two more losses to make the

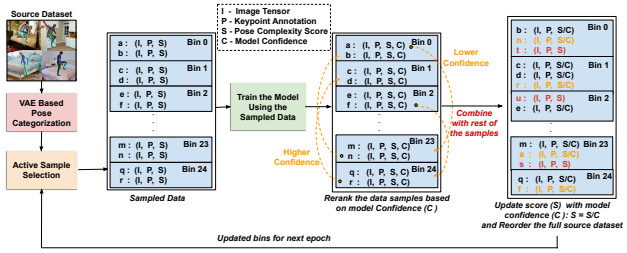

Figure 6. Active Curriculum Sampling organizes the source dataset into bins using active sampling for model training. Model predictions guide low confidence samples to higher bins, providing increased exposure. Updated scores and dataset reordering occur for the next epoch.

model aware of skeletal structure and to bring the features between the two domains closer. These losses are explained in the subsequent subsections. We refer to the total loss used by UDAPE [17] as L_{UDAPE}. This is a combination of the supervised loss and the consistency loss.

3.3. EvalPose Loss

The mean squared error loss has been conventionally used to train models for pose estimation [33]. However, this loss only considers the locations of the individual keypoints but fails to explicitly capture the relationship between the keypoints within a pose. We therefore, explicitly define keypoint pairs and a vector joining each pair to define the relationship between them. We compute the cosine similarity between the predicted keypoint vectors and the labeled keypoint vectors for the source domain, or, the pseudo-labeled keypoint vectors, by the teacher model, for the target domain. For a (pseudo-)labeled keypoint vector defined as $\mathbf{v}_d^k$ and a predicted keypoint vector defined as $\hat{\mathbf{v}}_\mathbf{d}^\mathbf{k}$, the loss is computed as follows:

$$\mathcal{L}_{\text{evalpose}} = -\sum_{k \in L} \frac{\mathbf{v}_d^k \cdot \hat{\mathbf{v}}_d^k}{\|\mathbf{v}_d^k\| \|\hat{\mathbf{v}}_d^k\|}, \quad d \in \{s, t\}. \tag{5}$$

Here L refers to set of keypoint pairs which define the vector connections and d refers to the source s or target t domain. Note that this loss is computed only within the same domain and not across domains. In Figure 8, we demonstrate that EvalPose is more sensitive to skeletal coherence despite minor changes in PCK, highlighting its effectiveness in ACTUPose.

3.4. Cross-Domain Loss

We additionally introduce a cross-domain feature loss to bring features of the source and target domain closer. For a batch of source and target domain samples we predict the pose from the student and the teacher models respectively. We find similar pairs of poses using a mean angular difference between poses. The angles are formed by custom

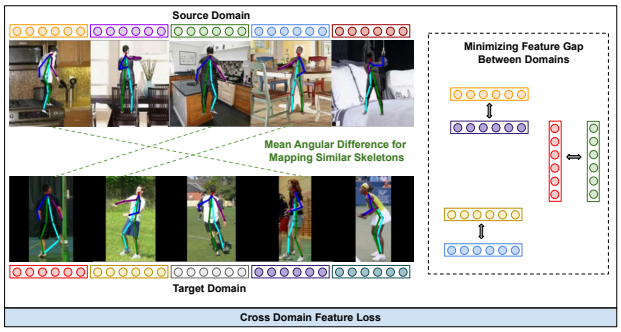

Figure 7. Mean angular difference is utilized to mine similar pairs in the source and target domain. The cross domain feature loss is applied on these pairs to improve the model's domain adaptation capabilities.

defined keypoint triplets within a pose. We prune out keypoints with low prediction confidence in the target domain data and their corresponding triplets before finding similar poses to avoid erroneous matches. Once the similar pairs are identified across the batches the student model is trained using the loss function defined below:

$$L_{\text{cross-domain}} = -\sum_{(z_1, z_2) \in Z} \frac{\mathbf{f}_s^{z_1} \cdot \mathbf{f}_t^{z_2}}{\|\mathbf{f}_s^{z_1}\| \|\mathbf{f}_t^{z_2}\|}, \tag{6}$$

where, Z is the set of similar pairs across source and target and $\mathbf{f}_t, \mathbf{f}_s$ represent the features for the target and source datapoint from the teacher and the student model respectively. The set Z is populated as follows. We define the set of angles Θ_s formed by the source domain predicted poses $\Theta_s = \{\theta_s^1 \dots \theta_s^S\}$ and similarly $\Theta_t = \{\theta_t^1 \dots \theta_t^T\}$ for the target domain. Here, $\theta_{s/t}$ represents a vector of all the angles between custom defined triplets for a single predicted pose in the source or target domains, S and T represent the number of images in the source and target batches. The mean angular difference between every pair of source and target batch poses is defined as follows:

$$\Delta\Theta^{(m,n)} = \frac{\|\theta_s^m - \theta_t^n\|}{K} \quad \forall m \in S, \forall n \in T. \tag{7}$$

$\Delta\Theta$ is an SxT matrix. Only those pose pairs with a value lesser than a threshold value Γ ($\Delta\Theta < \Gamma$) are considered similar pairs within the set Z. Hence, the total loss to train the student model is,

$$L = w_1 L_{UDAPE} + w_2 L_{evalpose} + w_3 L_{\text{cross-domain}}. \tag{8}$$

The teacher model gets updated through an exponential moving average setup following Mean Teacher [28].

4. Experiments and Results

In this section, we present the experimental evaluation of our proposed method, ACTUPose, across two key tasks: (1)

human body pose estimation and (2) hand pose estimation. These evaluations are conducted using benchmark datasets relevant to each domain. We assess the performance of ACTUPose by comparing it against state-of-the-art (SOTA) baselines, highlighting its improvements. Additionally, we employ the EvalPose score, a heuristic metric that quantifies the structural accuracy of the predicted pose in relation to the ground truth. To further support our numerical findings, we provide qualitative results, presented in Figure 8.

Datasets. The **SURREAL** [30] dataset is a synthetically-generated dataset rendered from sequences of human motion capture data. The dataset has 6 million labeled frames of human body poses covering a wide variety of actions. **Leeds Sports Pose** [15] (LSP) is a real-world outdoor human pose dataset capturing individuals in a wide range of poses, including challenging scenarios with occlusions. Comprising 2000 images, it provides annotations for key human body joint locations, primarily gathered during sports activities. **Human3.6M** [13] (H3.6M) is a real-world video dataset for human body pose estimation that includes data of diverse indoor activities. It has a total of 3.6 million frames. We follow the training and evaluation splits defined in [20]. The dataset has 5 subjects (S1, S5, S6, S7, S8) for training and the remaining 2 subjects (S9, S11) for testing. This split is typically adopted to train and evaluate models for human pose estimation. **Rendered Hand Pose Dataset** [40] (RHD), is a synthetic dataset for the task of hand pose estimation. It encompasses a wide range of hand poses captured under varying lighting conditions and comprises 41.2k training images, 2.7k test images, and annotations for 21 hand keypoints. **Hand-3D-Studio dataset** [39] (H3D), abbreviated as H3D, is a real-world dataset capturing multi-view indoor hand poses. It has a collection of 22k frames. Following a similar partitioning approach as used in the RegDA [14] framework, a subset of 3.2k frames is designated as the test set.

Evaluation Metrics. In this paper, we utilize PCK for quantitative analysis and EvalPose as a heuristic score to assess structural coherence and perceptual correctness. **PCK Score:** The Percentage of Correct Keypoints (PCK) measures the precision of body joint localization. A predicted joint is considered correct if its distance from the ground-truth location is within a specified threshold. We report results using PCK@0.05, which quantifies the proportion of correct predictions within 5% of the image size—higher values indicate greater accuracy. In addition to PCK, we use **EvalPose**, a visual perception score designed to evaluate the structural plausibility and perceptual realism of predicted poses. Unlike PCK, which focuses on numerical correctness, EvalPose captures the geometric consistency of poses, bridging the gap between keypoint accuracy and human perception. It follows the same formulation detailed in Section 3.3 and measures the structural similarity of predicted poses. As illustrated in figure 8, PCK alone may not fully capture variations in pose realism. EvalPose enhances interpretability by providing insights into the model's ability to generate natural and coherent poses, making it a valuable complementary measure.

Table 1. PCK@0.05 score on task SURREAL → LSP. Sld: shoulder, Elb: Elbow. We observe that our method ACTUPose outperforms the UDAPE model significantly across all joints.

Method	PCK						
	Sld	Elb	Wrist	Hip	Knee	Ankle	All
Source Only	51.5	65.0	62.9	68.0	68.7	67.4	63.9
Oracle	95.3	91.8	86.9	95.6	94.1	93.6	92.9
RegDA [14]	62.7	76.7	71.1	81.0	80.3	75.3	74.6
PoseDA [34]	**82.3**	78.4	73.2	74.8	79.7	78.7	77.9
UDAPE [17]	69.2	84.9	83.3	85.5	84.7	84.3	82.0
UDAPE + VAE-HM [9]	68.5	86.2	84.7	84.8	85.8	85.6	82.6
ACTUPose (Ours)	71.6	**87.7**	**86.5**	**88.8**	**87.5**	**87.1**	**84.9**

Table 2. Avg PCK@0.05 on benchmark tasks, SURREAL → Leed Sports Pose (LSP), SURREAL → Human3.6M (H3.6M) and Rendered Hand Pose (RHD) → Hand-3D-Studio (H3D).

Method	SURREAL → LSP	SURREAL → H3.6M	RHD → H3D
	AvgPCK	AvgPCK	AvgPCK
Source Only	63.9	67.3	61.8
Oracle	92.9	92.9	95.8
RegDA [14]	74.6	75.6	72.5
PoseDA [34]	77.9	79.6	N/A
UDAPE [17]	82.0	79.0	79.6
UDAPE + VAE-HM [9]	82.6	78.3	79.8
ACTUPose (Ours)	**84.9**	**82.8**	**80.9**

Experimental Setup Our framework builds upon UDAPE [17]. In the **training** phase of our experiment, the VAE-based (3.1.1) pose categorization orders samples into $N_b = 25$ bins arranged in increasing order of pose complexity. Employing the curriculum training strategy, we select samples across bins for each epoch, actively being shuffled on basis of model performance. This sampling and shuffling operation occurs before each epoch. The pacing parameter λ, governs the rate of progression through bins; a higher λ results in fewer epochs with a bin as the peak, while a decrease allows more time to sample from a bin per epoch. For the **Validation** phase, we categorize the target dataset into 25 bins of equal sizes using the category-based sampling approach outlined in Section 3.1.1. This categorization is useful to highlight our insights into the domain adaptation model's generalization capabilities across increasing pose complexities as there is a clear trend in model performance on increasingly complex bins.

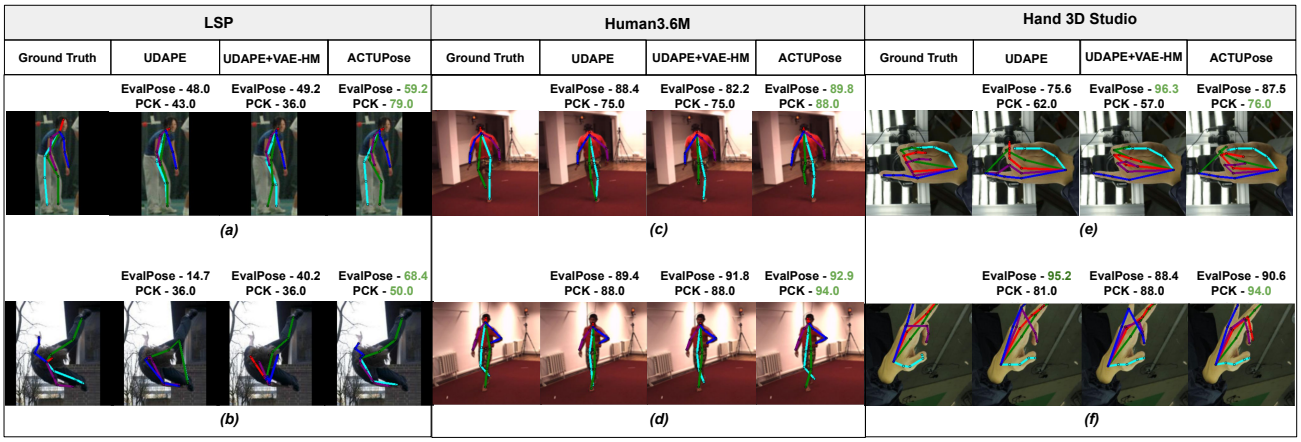

Figure 8. **Qualitative comparison** of ACTUPose, UDAPE[17], and UDAPE+VAE-HM[9] on LSP[15], Human3.6M[13], and Hand 3D Studio[39] datasets. ACTUPose consistently performs better, especially on challenging poses. While PCK measures keypoint localization, EvalPose is more sensitive to skeletal coherence. Significant EvalPose differences in samples like (c) and (f), despite minor PCK changes, highlight its effectiveness. This validates EvalPose as the loss function in ACTUPose, ensuring consistent improvements.

Quantitative Results. We evaluate our approach using well-established UDA benchmark experiments: (1) SURREAL → LSP, (2) SURREAL → H3.6M, and (3) RHD → H3D. The results of these experiments are summarized in Table 1 and Table 2. We compare our method against several baselines: **(a) Source-only training:** The model is trained only on the labeled source domain **(b) Oracle:** The model is trained directly on the target domain's training set. **(c) RegDA [14]:** A baseline UDA approach using adversarial domain adaptation. **(d) PoseDA [34]:** A baseline UDA approach leveraging hierarchical keypoints feature alignment to improve cross-domain pose estimation.**(e) UDAPE [17]:** A baseline UDA approach that follows a student-teacher learning paradigm. **(f) UDAPE + VAE-HM [9]**: Using a VAE-based categorization, we divide poses into 25 bins and select the most complex 8 bins for training the model. (Section 3.1.1). **(g) ACTUPose:** Our proposed approach, which incorporates an active curriculum training strategy. Table 1 presents the per-joint PCK scores and the overall average PCK for the SURREAL → LSP experiment. Our proposed method, ACTUPose, achieves superior performance compared to state-of-the-art UDA approaches, RegDA [14] and UDAPE [17], demonstrating consistent improvements across all joints as well as in overall accuracy. These results highlight the effectiveness of our training strategy in efficiently utilizing source domain samples to enhance adaptation to the target domain. Additionally, we observe that UDAPE + VAE-HM outperforms the UDAPE baseline while using only 33% of the source dataset, with its advantage being most pronounced in challenging scenarios, particularly within the hard bins (Figure 9). However, ACTUPose consistently achieves superior performance across all bins except the last three, ultimately

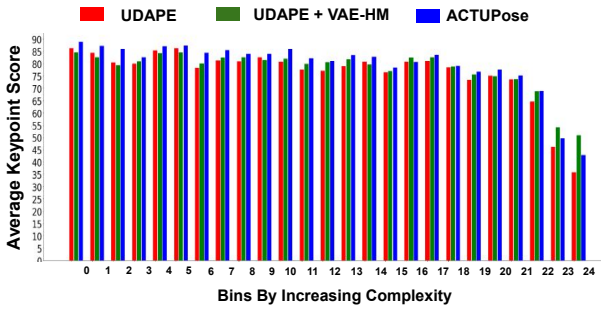

Figure 9. Comparison of UDAPE, UDAPE+VAE-HM, and ACTUPose on the SURREAL → LSP task, with bins arranged by increasing pose complexity. UDAPE+VAE-HM demonstrates superior performance in later bins due to its training on diverse and challenging poses. ACTUPose consistently outperforms both UDAPE and UDAPE+VAE-HM across nearly all bins, highlighting its effectiveness in domain adaptation.

surpassing both methods overall. Table 2 reports the PCK scores for all benchmark experiments, further demonstrating the effectiveness of ACTUPose. Our approach consistently surpasses state-of-the-art models across all datasets, achieving a 2.3% improvement in SURREAL → LSP, 4.5% in SURREAL → H3.6M, and 1.1% in RHD → H3D. These results underscore the robustness of ACTUPose in enhancing domain adaptation performance across diverse datasets.

Figure 9 illustrates the performance of different methods across 25 clustered bins of the validation set. Each bin contains three bar plots: the first represents UDAPE[17], the second corresponds to the UDAPE + VAE-HM [9] for domain adaptation, and the third represents the proposed method, ACTUPose. To maintain visual interpretability, other methods are omitted. The y-axis denotes the average

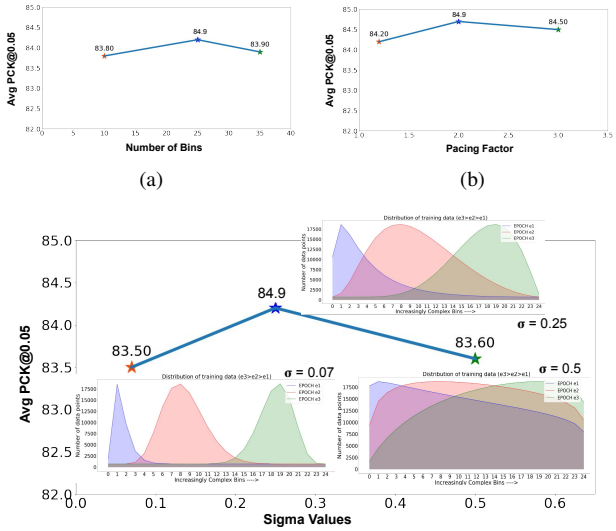

Figure 10. Ablation Study showing the effect of different training parameters on Avg PCK@0.05. (a) Effect of number of bins (b) Effect of the pacing factor and (c) Impact of sigma value on data distribution. Best optimal performance is obtained at $N_b = 25$, $\lambda = 2.0$ and $\sigma = 0.25$

keypoint score, calculated as the mean accuracy per pose within each bin. As bin complexity increases, the advantage of UDAPE + VAE-HM becomes more evident, benefiting from training on diverse and high-variation poses captured in the later bins. However, ACTUPose follows a progressive adaptation strategy, transitioning from simple to complex bins, leading to overall performance gains across all bins and consistently surpassing state-of-the-art models.

Qualitative Results. Figure 8 presents a qualitative comparison of ACTUPose against UDAPE and UDAPE + VAE-HM across three benchmark UDA tasks: human pose estimation on LSP and Human3.6M datasets, and hand pose estimation on the H3D dataset. PCK is used to assess keypoint localization accuracy, while EvalPose evaluates the structural plausibility of predicted poses. ACTUPose consistently outperforms baseline methods, achieving higher PCK scores across all samples. Moreover, EvalPose captures structural refinements even in cases where PCK shows minimal variation, as seen in samples (c) and (e), highlighting ACTUPose's ability to produce anatomically coherent poses. This further validates the use of EvalPose as a loss function within ACTUPose, ensuring consistent improvements across all samples.

Ablation Study. We conduct an ablation study to examine the effect of three key parameters in Active Curriculum Learning: the number of bins, the pacing factor (λ), and the imapct of (σ) value on data distribution. **Number of Bins:** The number of bins determines how the dataset is divided into bins using the reconstruction error from Pose-VAE 3.1.1. Too few bins mix different complexity levels, while too many bins create small groups with very limited data. The Avg PCK@0.05 for the SURREAL $\rightarrow$ LSP dataset with respect to change in the number of bins is shown in figure 10(a), From figure, We observe that the best performance is achieved at 25 bins. **Pacing Factor:** The pacing factor λ determines the rate at which the Gaussian curve's peak moves towards the harder bins. A very high λ value can end up spending very less time or epochs on less complex data leading to not learning the basic poses, while a very low λ value slows the training by overspending time on early epochs and increases the risk of overfitting. The Avg PCK@0.05 for the SURREAL $\rightarrow$ LSP dataset at different λ value is shown in figure 10(b). From figure, We observe that the optimal performance is achieved at $\lambda = 2$. **Impact of Sigma:** The parameter σ controls how much data from each bin contributes compared to the peak bin at any given epoch. A larger σ results in a broad curve, sampling data from distant bins, which disrupts the curriculum and includes most of the dataset in every epoch. A very small σ creates a narrow curve, sampling only from nearby bins. Figure 10(c) shows Avg PCK@0.05 for the SURREAL $\rightarrow$ LSP dataset across σ values. We find that $\sigma = 0.25$ provides a well-structured curriculum and optimal model performance.

Additionally, we conducted an extensive ablation study on the weights of different loss functions used in the proposed method, ACTUPose. Keeping $w_1 = 1.0$ (UDAPE loss) fixed, we varied w_2 (EvalPose loss) and w_3 (Cross-domain loss) within the range of 0 to 1. Our experiments show that incorporating both losses improves performance, with the best results achieved when $w_2 = 0.1$ and $w_3 = 0.01$, yielding the highest Avg PCK of **84.9** on the SURREAL $\rightarrow$ LSP, **82.8** on the SURREAL $\rightarrow$ Human3.6M and **80.9** on the RHD $\rightarrow$ H3D.

5. Conclusion

In conclusion, our work introduces ACTUPose, a novel unsupervised domain adaptation framework for pose estimation that enhances generalization across diverse target domains. By leveraging an active curriculum training strategy, ACTUPose progressively adapts to increasing pose complexity, ensuring a more structured learning process. The introduction of a structural consistency loss refines skeletal predictions, while a cross-domain feature alignment mechanism optimally leverages unlabeled real data. Extensive evaluations on multiple benchmark datasets demonstrate that ACTUPose consistently outperforms existing methods, achieving state-of-the-art accuracy and robustness in the pose estimation task.

References

[1] Roberto Alcover-Couso, Juan C. Sanmiguel, Marcos Escudero-Viñolo, and Pablo Carballeira. Per-class curriculum for unsupervised domain adaptation in semantic segmentation. *Vis. Comput.*, 41:901–919, 2024. 2

[2] Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. Curriculum learning. In *Proceedings of the 26th annual international conference on machine learning*, pages 41–48, 2009. 2

[3] Olivier Chapelle, Bernhard Scholkopf, and Alexander Zien. Semi-supervised learning (chapelle, o. et al., eds.; 2006)[book reviews]. *IEEE Transactions on Neural Networks*, 20(3):542–542, 2009. 2

[4] Dongjie Chen, Kartik Patwari, Zhengfeng Lai, Sen-Ching Samson Cheung, and Chen-Nee Chuah. Empowering source-free domain adaptation with mllm-driven curriculum learning. *ArXiv*, abs/2405.18376, 2024. 2

[5] Yilun Chen, Zhicheng Wang, Yuxiang Peng, Zhiqiang Zhang, Gang Yu, and Jian Sun. Cascaded pyramid network for multi-person pose estimation. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7103–7112, 2018. 2

[6] Jaehyun Choi, Junwon Ko, Dong-Jae Lee, and Junmo Kim. Ah-ocda: Amplitude-based curriculum learning and hopfield segmentation model for open compound domain adaptation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2025. 2

[7] Yan Dai, Beitao Chen, Lianli Gao, Jingkuan Song, and Heng Tao Shen. Dmh-cl: Dynamic model hardness based curriculum learning for complex pose estimation. *IEEE Transactions on Multimedia*, pages 1–14, 2023. 2

[8] Yann Desmarais, Denis Mottet, Pierre Slangen, and Philippe Montesinos. A review of 3d human pose estimation algorithms for markerless motion capture. *Computer Vision and Image Understanding*, 212:103–275, 2021. 1

[9] Isha Dua, Arjun Sharma, Shuaib Ahmed, and Rahul Tallamraju. Towards effective synthetic data sampling for domain adaptive pose estimation. In *NeurIPS 2023 Workshop on Synthetic Data Generation with Generative AI*, 2023. 2, 6, 7

[10] Chen Gong, Dacheng Tao, Stephen J Maybank, Wei Liu, Guoliang Kang, and Jie Yang. Multi-modal curriculum learning for semi-supervised image classification. *IEEE Transactions on Image Processing*, 25(7):3249–3260, 2016. 2

[11] Alex Graves, Marc G Bellemare, Jacob Menick, Remi Munos, and Koray Kavukcuoglu. Automated curriculum learning for neural networks. pages 1311–1320, 2017. 2

[12] Jihye Hwang, John Yang, and Nojun Kwak. Exploring rare pose in human pose estimation. *IEEE Access*, 8:194964–194977, 2020. 2

[13] 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*, 36:1325–1339, 2014. 2, 6, 7

[14] Junguang Jiang, Yifei Ji, Ximei Wang, Yufeng Liu, Jianmin Wang, and Mingsheng Long. Regressive domain adaptation for unsupervised keypoint detection. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6776–6785, 2021. 6, 7

[15] Sam Johnson and Mark Everingham. Clustered pose and nonlinear appearance models for human pose estimation. In *British Machine Vision Conference*, 2010. 1, 2, 6, 7

[16] Nazmul Karim, Niluthpol Chowdhury Mithun, Abhinav Rajvanshi, Han-Pang Chiu, Supun Samarasekera, and Nazanin Rahnavard. C-sfda: A curriculum learning aided self-training framework for efficient source free domain adaptation. *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 2

[17] Donghyun Kim, Kaihong Wang, Kate Saenko, Margrit Betke, and Stan Sclaroff. A unified framework for domain adaptive pose estimation. *ArXiv*, abs/2204.00172, 2022. 1, 2, 4, 5, 6, 7

[18] Seunghyeon Kim, Jaehoon Choi, Taekyung Kim, and Changick Kim. Self-training and adversarial background regularization for unsupervised domain adaptive one-stage object detection. pages 6092–6101, 2019. 2

[19] M Kumar, Benjamin Packer, and Daphne Koller. Self-paced learning for latent variable models. *Advances in neural information processing systems*, 23, 2010. 2

[20] Sijin Li and Antoni B. Chan. 3d human pose estimation from monocular images with deep convolutional neural network. In *Asian Conference on Computer Vision*, 2014. 6

[21] Siyang Li, Xiangxin Zhu, Qin Huang, Hao Xu, and C-C Jay Kuo. Multiple instance curriculum learning for weakly supervised object detection. *arXiv preprint arXiv:1711.09191*, 2017. 2

[22] David McClosky, Eugene Charniak, and Mark Johnson. Effective self-training for parsing. pages 152–159, 2006. 2

[23] Dong-Hyun Lee Pseudo-Label. The simple and efficient semi-supervised learning method for deep neural networks. pages 1–6, 2013. 2

[24] Stephan R. Richter, Vibhav Vineet, Stefan Roth, and Vladlen Koltun. Playing for data: Ground truth from computer games. In *Computer Vision – ECCV 2016*, pages 102–118. Springer, 2016. 1

[25] Grégory Rogez and Cordelia Schmid. Mocap-guided data augmentation for 3d pose estimation in the wild. *Advances in neural information processing systems*, 29, 2016. 2

[26] Aruni RoyChowdhury, Prithvijit Chakrabarty, Ashish Singh, SouYoung Jin, Huaizu Jiang, Liangliang Cao, and Erik Learned-Miller. Automatic adaptation of object detectors to new domains using self-training. pages 780–790, 2019. 2

[27] Petru Soviany, Claudiu Ardei, Radu Tudor Ionescu, and Marius Leordeanu. Image difficulty curriculum for generative adversarial networks (cugan). pages 3463–3472, 2020. 2

[28] Antti Tarvainen and Harri Valpola. Weight-averaged consistency targets improve semi-supervised deep learning results. *ArXiv*, abs/1703.01780, 2017. 5

[29] Radu Tudor Ionescu, Bogdan Alexe, Marius Leordeanu, Marius Popescu, Dim P Papadopoulos, and Vittorio Ferrari. How hard can it be? estimating the difficulty of visual search in an image. pages 2157–2166, 2016. 2

[30] Gül Varol, Javier Romero, Xavier Martin, Naureen Mahmood, Michael J. Black, Ivan Laptev, and Cordelia Schmid. Learning from synthetic humans. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4627–4635, 2017. 1, 2, 6

[31] Cheng Wang, Qian Zhang, Chang Huang, Wenyu Liu, and Xinggang Wang. Mancs: A multi-task attentional network with curriculum sampling for person re-identification. pages 365–381, 2018. 2

[32] Jiasi Wang, Xinggang Wang, and Wenyu Liu. Weakly-and semi-supervised faster r-cnn with curriculum learning. pages 2416–2421, 2018. 2

[33] Bin Xiao, Haiping Wu, and Yichen Wei. Simple baselines for human pose estimation and tracking. *ArXiv*, abs/1804.06208, 2018. 4, 5

[34] Jie Xu, Yunan Liu, Jian Yang, and Shanshan Zhang. Hierarchical keypoints feature alignment for domain adaptive pose estimation. *Neurocomputing*, 2024. 6, 7

[35] Jun-Sang Yoo, Jung, and Seung-Won. Survey on in-vehicle datasets for human pose estimation. In *2022 International Conference on Electronics, Information, and Communication (ICEIC)*, pages 1–2, 2022. 1

[36] Andrei Zanfir, Mihai Zanfir, Alex Gorban, Jingwei Ji, Yin Zhou, Dragomir Anguelov, and Cristian Sminchisescu. Hum3dil: Semi-supervised multi-modal 3d humanpose estimation for autonomous driving. In *Proceedings of The 6th Conference on Robot Learning, PMLR*, pages 1114–1124, 2022. 1

[37] Dingwen Zhang, Junwei Han, Long Zhao, and Deyu Meng. Leveraging prior-knowledge for weakly supervised object detection under a collaborative self-paced curriculum learning framework. *International Journal of Computer Vision*, 127:363–380, 2019. 2

[38] Feng Zhang, Xiatian Zhu, and Chen Wang. Comprehensive survey on single-person pose estimation in social robotics. *Int J of Soc Robotics*, 14:1995–2008, 2022. 1

[39] Zheng Fa Zhao, Tianyao Wang, Siyu Xia, and Yangang Wang. Hand-3d-studio: A new multi-view system for 3d hand reconstruction. *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2478–2482, 2020. 2, 6, 7

[40] Christiane Zimmermann and Thomas Brox. Learning to estimate 3d hand pose from single rgb images. *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 4913–4921, 2017. 6

[41] Yang Zou, Zhiding Yu, BVK Kumar, and Jinsong Wang. Unsupervised domain adaptation for semantic segmentation via class-balanced self-training. pages 289–305, 2018. 2

[42] Yang Zou, Zhiding Yu, Xiaofeng Liu, BVK Kumar, and Jinsong Wang. Confidence regularized self-training. pages 5982–5991, 2019. 2