CLIP2Pose: Frozen CLIP as Semantic Guide for Domain Adaptive Pose Estimation

Jiawen Li[1], **Fei Jiang**[2*], **Dandan Zhu**[3], **Jinxin Shi**[1], **Aimin Zhou**[1]

[1]Shanghai Institute of Artificial Intelligence for Education, East China Normal University, Shanghai, China
[2]Cognitive Intelligence Lab, Chongqing Academy of Science and Technology, Chongqing, China
[3]School of Computer Science and Technology, East China Normal University, Shanghai, China

Abstract

Unsupervised domain adaptive pose estimation is a fundamental yet challenging task due to the need to transfer from labeled synthetic data to unlabeled real data. Nevertheless, the underlying pose semantics, which are governed by spatial structure, remain largely consistent across domains. This observation motivates the use of vision-language models, which provide domain-invariant representations that align well with high-level semantic concepts. Motivated by this, we propose CLIP2Pose, a novel framework that leverages the semantic robustness of frozen CLIP encoders to facilitate cross-domain generalization. We first introduce a semantic-driven prompt mechanism that encodes structural priors, domain-specific appearance, and instance-level context into the image representation. This guides the model to focus on semantically meaningful and structurally relevant features. Next, we propose a semantic modulation module that adaptively refines visual features by conditioning them on prompt derived embeddings, enhancing alignment between semantics and visual patterns. To further bridge the modality and domain gaps, we design a directional alignment loss that encourages consistent structural reasoning across both vision and language representations. Extensive experiments on domain adaptive human body and hand pose benchmarks show that CLIP2Pose achieves state-of-the-art performance.

Introduction

Pose estimation is a fundamental task in computer vision, supporting a wide range of applications including gesture recognition, human-computer interaction, sign language interpretation, and augmented reality (Doosti et al. 2020; Cao et al. 2021). Despite recent progress, advancement in this task is often limited by the scarcity of large-scale annotated datasets, as manual labeling is both costly and error-prone. To mitigate this, synthetic datasets offer an alternative with automatically generated and structurally consistent annotations (Zimmermann and Brox 2017; Yuan et al. 2018). However, models trained solely on synthetic data often suffer significant performance drops when applied to real-world scenarios due to the substantial domain gap (Li et al. 2023).

Unsupervised Domain Adaptation (UDA) aims to bridge this gap by transferring knowledge from labeled synthetic domain (source) to unlabeled real domain (target) (Jiang et al. 2021; Kim et al. 2022; Peng, Zheng, and Chen 2023; Lin, Yang, and Yao 2023; Ohkawa et al. 2022). Most recent UDA methods for pose estimation focus on feature-level alignment through adversarial learning (Jiang et al. 2021; Kim et al. 2022), domain-specific disentanglement (Peng, Zheng, and Chen 2023), or appearance-level adaptation (Lin, Yang, and Yao 2023; Ohkawa et al. 2022). While recent approaches have made notable progress, domain shifts in articulation, scale, and visual context continue to disrupt the structural consistency of poses and limit the generalization ability of current methods.

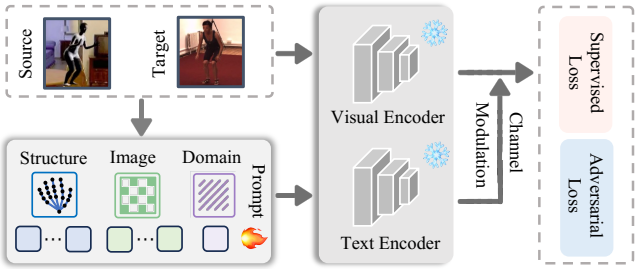

Figure 1: Overview of the proposed CLIP2Pose framework. A frozen CLIP model is used to align source and target domains via domain-invariant semantics. To adapt CLIP for pose estimation, we introduce learnable textual prompts comprising structure, domain, and image components.

Vision-Language Models (VLMs), such as CLIP (Radford et al. 2021), provide a powerful alternative for domain adaptation by aligning images and text within a shared semantic space, thereby capturing domain-invariant information beyond low-level appearance. Recent advances have explored prompt learning as a lightweight adaptation strategy, introducing learnable tokens that encode task or domain specific knowledge without requiring full model finetuning (Zhao et al. 2025). These approaches have demonstrated remarkable performance in cross domain classification (Ge et al. 2023) and segmentation (Wu et al. 2024), highlighting the potential of semantic level alignment for effective domain adaptation.

However, the extension of vision-language models to continuous, structured outputs remains underexplored. Un-

*Corresponding Author

like classification tasks, pose estimation inherently contains strong structural priors, such as the kinematic relationships between joints (Kumarlal Fernando 1975). These priors are consistent across domains and are crucial for accurate prediction. Existing prompt-based techniques are primarily designed for discrete label spaces and lack the expressiveness to capture such spatial dependencies. This discrepancy introduces two key challenges for adapting CLIP-like models to pose estimation: (1) *How can structural priors of the hand be effectively expressed in the textual prompt space to help CLIP comprehend the requirements of keypoint detection tasks?* (2) *How can the model learn domain invariant, structure aware representations that generalize across varied poses and visual conditions?*

To address these problems, we propose CLIP2Pose, a novel framework that leverages the domain-invariant semantics of frozen CLIP model to enhance UDA pose estimation. An overview of the proposed framework is illustrated in Figure 1. CLIP2Pose introduces language-driven prompts as semantic anchors, guiding the visual encoder toward structure aware and domain-agnostic representations. We first design a Semantic-Driven Prompt (SDP) mechanism that enriches image encoding through three complementary perspectives: structural topology, domain style, and instance cues. Among them, structure Tokens encode the intrinsic kinematic layout of the body or hand via a graph-based formulation. This design injects spatial inductive bias into the learning process and reinforces pose consistency across domains. However, the alignment between visual features and structural semantics cannot rely solely on spatial priors. Visual encoders often overemphasize appearance sensitive channels, such as color, texture, and local edge distributions, which vary significantly across domains. To address this limitation, we introduce a Semantic Modulation of Visual Features (SMVF) module that modulates channel-wise responses under the guidance of prompt derived embeddings. SMVF enhances the saliency of task relevant structures while attenuating domain variant noise, thereby improving the robustness of keypoint localization in unseen domains. To further consolidate semantic consistency, we impose a Directional Alignment Loss (DAL), which encourages the global visual representation to align with the prompt embedding within a shared semantic space. By minimizing the angular deviation between modalities, this constraint ensures that the structural semantics encoded by language are consistently preserved in visual predictions, irrespective of domain shifts. The main contributions of this paper are as follows:

- We propose CLIP2Pose, a novel framework for UDA pose estimation. It leverages frozen CLIP models to provide language-informed structural priors that guide cross-domain representation learning.
- We design a semantic-guided adaptation strategy that injects structured language priors into the visual pipeline, enabling the model to dynamically align, modulate, and regularize visual features with respect to domain-invariant semantics.
- Extensive experiments demonstrate that CLIP2Pose achieves state-of-the-art performance across multiple domain adaptive pose estimation benchmarks, including both hand and full-body settings.

Realated Work

Domain Adaptive Pose Estimation

Early efforts in domain adaptive pose estimation, such as CC-SSL (Mu et al. 2020), RegDA (Jiang et al. 2021), and TransPar (Han, Sun, and Yin 2022), primarily focused on aligning feature distributions across domains through unified architectures or shared backbone encoders. Among them, adversarial learning emerged as a key strategy, offering a principled way to promote domain invariance through distribution-level alignment. Representative methods such as RegDA (Jiang et al. 2021) and MarsDA (Jin et al. 2022) explicitly incorporate domain discriminators into the learning process, enabling the model to align source and target feature distributions while preserving task-specific semantics. Beyond feature-level alignment, recent approaches have extended adversarial learning to other modalities, including pixel-level appearance, structural dependency modeling (Zhang et al. 2020), and source-free adaptation (Peng, Zheng, and Chen 2023). In addition, some recent works explore attention-based adversarial alignment (Deng et al. 2024), where adaptive discriminators are guided by keypoint-level confidence to focus on critical joint regions. Motivated by this, our approach also incorporates adversarial learning to enforce structure aware domain alignment, enabling robust keypoint localization across domains without relying on unreliable pseudo-labels.

Vision Language Models

Vision-Language Models (VLMs) have demonstrated strong capabilities in learning cross-modal correspondences by leveraging large-scale image-text pairs during pretraining. These models are typically trained under a variety of supervision signals, such as masked language modeling (Kim, Son, and Kim 2021), image-text matching (Huang et al. 2021), and contrastive learning (Jia et al. 2021; Zhang et al. 2022a; Chen et al. 2021). These models have demonstrated strong performance in open-vocabulary and few-shot visual recognition. However, adapting their pretrained knowledge to downstream tasks under domain shift remains challenging. To mitigate this, recent studies have focused on enhancing their transferability. One common approach is to introduce feature adapters, which serve as lightweight modules to bridge the gap between task-specific objectives and pretrained representations (Gao et al. 2024; Zhang et al. 2023; Bai et al. 2024b). In parallel, attention-based mechanisms (Guo et al. 2023) and memory-augmented modules (Zhang et al. 2022b) have been utilized to enhance context aggregation and feature reuse. While previous efforts have shown promise, adapting VLMs to structured prediction problems like pose estimation remains underexplored. Given the strong generalization ability of VLMs, we are motivated to investigate their potential in this setting.

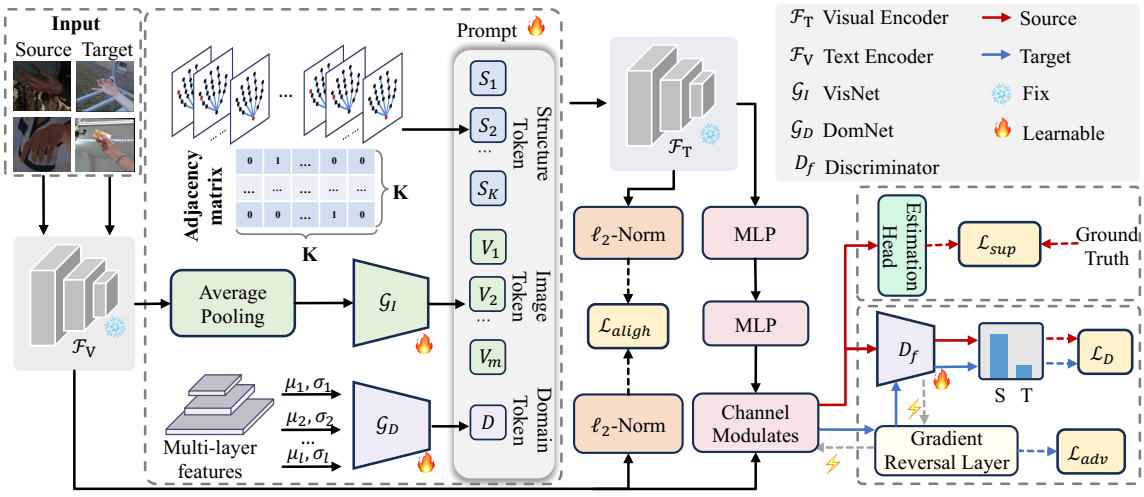

Figure 2: Framework of the proposed CLIP2Pose method for unsupervised domain adaptation in pose estimation. The model consists of frozen CLIP encoders and a learnable prompt module composed of structure, domain, and image components. These prompts guide the visual features through semantic alignment and feature modulation. A directional alignment loss further enforces consistency between visual and textual semantics, enabling robust keypoint prediction across domains.

Vision-Language Models in Domain Adaptation

Recent studies have shown that vision-language models, particularly CLIP, hold strong potential for unsupervised domain adaptation (UDA) due to their robust visual-textual representations. Early works such as PADCLIP (Lai et al. 2023) and AD-CLIP (Singha et al. 2023) demonstrated that encoding domain and class semantics into learnable prompts enables effective cross-domain generalization without modifying the backbone. Building on this idea, DAPrompt (Ge et al. 2023) introduced dynamic prompt templates to better adapt to domain-specific distributions, while CLIP-Div (Zhu, Chen, and Wang 2024) employed divergence-based losses guided by text to align feature distributions. To further enhance generalization, ProGrad (Zhu et al. 2023) mitigates overfitting by aligning prompt updates with pretrained gradients, and MaPLe (Khattak et al. 2023) jointly optimizes visual and textual prompts to strengthen modality interactions. PDA (Bai et al. 2024a) introduces a method that integrates domain knowledge into prompt learning using a two-branch prompt tuning framework with base and alignment branches. ReCLIP (Hu et al. 2024) addresses misalignment by refining the shared embedding space through pseudo-label-driven self-training. Inspired by these advances, we design task-specific prompts that not only encode semantic priors but also guide visual representation learning under domain shifts.

Method

Overview

Given a labeled source domain $\mathcal{S} = \{(x_i^s, y_i^s)\}_{i=1}^{N_s}$ and an unlabeled target domain $\mathcal{T} = \{x_i^t\}_{i=1}^{N_t}$, our goal is to predict a set of K keypoints in the form of Gaussian heatmaps $H \in \mathbb{R}^{H \times W \times K}$. A key challenge in this setting is the domain gap between $\mathcal{S}$ and $\mathcal{T}$. Our method CLIP2Pose leverages language-driven semantic guidance to enhance cross-domain generalization, as illustrated in Figure 2. A prompt $\mathbf{P}(x)$ is constructed per image, capturing structural, stylistic, and instance-specific cues. It is encoded into a semantic embedding $e = \mathcal{F}_t(\mathbf{P}(x))$ using a frozen CLIP text encoder, while the visual encoder $\mathcal{F}_v$ extracts features $z = \mathcal{F}_v(x)$. The semantic embedding e provides structural guidance to the visual features z through a modulation function $\mathcal{M}$, resulting in enhanced features $\tilde{z} = \mathcal{M}(z, e)$. To align the enhanced features $\tilde{z}$ with the semantic embedding, we apply a directional alignment loss $\mathcal{L}_{align}$. In addition, a domain discriminator is employed to encourage domain-invariant representations via adversarial training. The overall objective combines supervised loss on the source domain, adversarial loss for domain alignment, and the semantic alignment term:

$$\mathcal{L} = \mathcal{L}_{sup} + \mathcal{L}_{adv} + \lambda \cdot \mathcal{L}_{align}. \quad (1)$$

In the following, we detail the semantic prompt module.

Semantic-Driven Prompting

We observe two key properties that help mitigate domain discrepancies in pose estimation. First, the semantics of human motion are primarily defined by structural relationships, which remain consistent across domains. For example, a gesture like "index finger touching thumb" or "bending the knee" corresponds to invariant topological arrangements of keypoints. Second, domain gaps mainly stem from local, non-structural factors such as lighting or texture, which do not alter the underlying spatial configuration. Based on these insights, we propose a Semantic-Driven Prompt (SDP) mechanism that injects structural priors, domain styles, and instance-specific cues into the image encoding process. Given an input image x, we construct a structured prompt sequence with three types of tokens:

$$\mathbf{P}(x) = [S]_1 \dots [S]_K [V]_1 \dots [V]_M [D] \in \mathbb{R}^{T \times D}, \quad (2)$$

where $[S]$ denotes structure tokens, each capturing the semantic relation of a specific keypoint. $[V]$ represents instance-specific tokens that encode discriminative local content within the image. $[D]$ is a domain-style token that captures the global visual appearance of the input. And T is the overall token length.

Structure Token The spatial topology of the pose skeleton remains consistent across domains, while appearance varies. To leverage this invariance, we introduce shared structure tokens that encode the pose skeleton as a graph and propagate relational information using Graph Convolutional Networks (GCNs). Formally, the skeleton is represented as an undirected graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where $\mathcal{V} = \{1, \ldots, K\}$ denotes keypoint nodes and $\mathcal{E} \subseteq \mathcal{V} \times \mathcal{V}$ the set of physical connections. The graph is encoded by an adjacency matrix $\mathbf{A} \in \{0,1\}^{K \times K}$, with $\mathbf{A}_{ij} = 1$ if $(i,j) \in \mathcal{E}$, and 0 otherwise. Adjacency matrix details are provided in the Appendix. Each node $i \in \mathcal{V}$ is initialized with a learnable embedding $\mathbf{s}_i^{(0)} \in \mathbb{R}^D$, forming the structure token matrix:

$$\mathbf{S}^{(0)} = [\mathbf{s}_1^{(0)}; \ldots; \mathbf{s}_K^{(0)}] \in \mathbb{R}^{K \times D}. \tag{3}$$

We then apply GCNs over the structure tokens using the normalized adjacency matrix with self-loops:

$$\tilde{\mathbf{A}} = \mathbf{A} + \mathbf{I}, \quad \tilde{\mathbf{D}}_{ii} = \sum_j \tilde{\mathbf{A}}_{ij}, \quad \hat{\mathbf{A}} = \tilde{\mathbf{D}}^{-1/2} \tilde{\mathbf{A}} \tilde{\mathbf{D}}^{-1/2}. \tag{4}$$

Structure tokens are updated as:

$$\mathbf{S}^{(1)} = \delta(\hat{\mathbf{A}} \mathbf{S}^{(0)} \mathbf{W}_1), \quad \ldots, \quad \mathbf{S}^{(n)} = \hat{\mathbf{A}} \mathbf{S}^{(n-1)} \mathbf{W}_n, \tag{5}$$

where $\mathbf{W}_i$ are trainable weights and δ is ReLU. The final output $[S] = \mathbf{S}^{(n)} \in \mathbb{R}^{K \times D}$ encodes structure tokens enriched with both local and global topology.

Image Token To model image-specific semantic variations such as occlusions, non-canonical poses, and background clutter, we introduce a set of learnable image tokens $\{\mathbf{v}_m\}_{m=1}^M$, where each $\mathbf{v}_m \in \mathbb{R}^D$ is initialized as a trainable parameter. These tokens are dynamically adapted to each input image based on its global semantic context. Given the visual feature map $z \in \mathbb{R}^{C \times H \times W}$, it is globally pooled and linearly projected to produce a semantic-aligned embedding $z_e \in \mathbb{R}^D$. z_e is passed through a lightweight two-layer bottleneck network $\mathcal{G}_I$, consisting of a linear layer, a ReLU activation, and another linear layer to produce an offset vector $\pi \in \mathbb{R}^D$. Each image token is then adapted as follows:

$$[V]_m = \mathbf{v}_m + \pi, \quad m = 1, \ldots, M, \tag{6}$$

where $[V]_m$ denotes the m-th instance-specific image token in the final prompt. Following the strategy of CoCoOp (Zhou et al. 2022), all tokens share the same offset π, allowing contextual adaptation while preserving token-level diversity through their unique base embeddings $\mathbf{v}_m$.

Domain Token To enhance the encoder's adaptability to domain-specific visual styles, we introduce a domain token that captures global stylistic cues from the input image. Given an image x, we extract multi-scale features $f_l|_{l=1}^4$ from different encoder stages and compute their mean μ_l and standard deviation σ_l to represent channel-wise statistics. These are concatenated into a global style vector $\mathbf{z}_g \in \mathbb{R}^C$ and projected into the prompt space using a two-layer MLP, DomNet $\mathcal{G}_D : \mathbb{R}^C \to \mathbb{R}^D$. The resulting domain token $[D] = \mathcal{G}_D(\mathbf{z}_g)$ is then inserted into the prompt sequence to guide the model in adapting to style shifts across domains.

Semantic Modulation of Visual Features

A key challenge in UDA pose estimation lies in the sensitivity of visual encoders to low-level appearance cues such as color, texture, and edge intensity, which often vary across domains. In contrast, language prompts encode high-level structural priors that are inherently domain-invariant, offering robust semantic guidance. To leverage this property, we propose a Semantic Modulation of Visual Features (SMVF) mechanism, which adjusts global visual activations using prompt derived semantics.

Given an input image $x \in \mathbb{R}^{3 \times H \times W}$, the visual encoder $\mathcal{F}_v$ extracts feature maps $z = \mathcal{F}_v(x) \in \mathbb{R}^{C \times H \times W}$. We apply global average pooling to obtain a compact visual vector $\bar{z} \in \mathbb{R}^C$. Simultaneously, the text encoder $\mathcal{F}_t$ processes the structured prompt $\mathbf{P}(x)$, yielding a semantic embedding $e \in \mathbb{R}^D$. To align the two modalities, we project the text embedding into the visual space via a two-layer MLP:

$$\gamma = \sigma\left(\mathbf{W}^{(2)}\, \delta\left(\mathbf{W}^{(1)} e\right)\right), \quad \gamma \in \mathbb{R}^C,$$

where δ and σ denote ReLU and Sigmoid activations, respectively. The modulation vector γ is then applied to the pooled visual feature $\bar{z}$ via element-wise multiplication to obtain a semantically guided representation:

$$\tilde{z} = \gamma \odot \bar{z}.$$

This is applied before estimation head and shared across samples without introducing domain-specific components.

Directional Alignment Loss

To promote semantic consistency across modalities and domains, we introduce a directional alignment loss that encourages global visual features to align with prompt derived semantic embeddings. Let the visual encoder $\mathcal{F}_v$ extract a feature map $z \in \mathbb{R}^{C \times H \times W}$ from an input image x. We apply global average pooling over spatial dimensions to obtain a compact visual representation $\bar{z} \in \mathbb{R}^D$. The text encoder $\mathcal{F}_t$ produces a structured semantic embedding $e \in \mathbb{R}^D$ from the input prompt. Since $\bar{z}$ and e lie in different spaces, we use a learnable projection matrix $\mathcal{W}_e \in \mathbb{R}^{C \times D}$ to map e into the visual feature space:

$$\tilde{e} = \mathcal{W}_e e \in \mathbb{R}^C.$$

To compare them in a geometry-aware manner, we normalize both vectors onto the unit hypersphere using ℓ_2 normalization. Let $\mathcal{P}(v) = \frac{v}{\|v\|_2}$ denote this projection. The directional alignment loss is then defined as the cosine distance between the normalized embeddings:

$$\mathcal{L}_{\text{align}} = 1 - \langle \mathcal{P}(\bar{z}),\ \mathcal{P}(\tilde{e}) \rangle\,. \tag{7}$$

Method	SURREAL→LSP							SURREAL→Human3.6M						
	Sld	Elb	Wrist	Hip	Knee	Ankle	All	Sld	Elb	Wrist	Hip	Knee	Ankle	All
Source Only	57.4	73.6	59.3	55.5	63.1	66.6	56.7	48.1	78.9	64.7	40.9	75.5	78.6	55.3
CC-SSL (Mu et al. 2020)	36.8	66.3	63.9	59.6	67.3	70.4	60.7	44.3	68.5	55.2	22.2	62.3	57.8	51.7
MDAM (Li and Lee 2021)	61.4	77.7	75.5	65.8	76.7	78.3	69.2	51.7	83.1	68.9	17.7	79.4	76.6	62.9
RegDA (Jiang et al. 2021)	62.7	76.7	71.1	81.0	80.3	75.3	74.6	73.3	86.4	72.8	54.8	82.0	84.4	75.6
Uniframe (Kim et al. 2022)	69.2	84.9	83.3	85.5	84.7	84.3	82.0	78.1	89.6	81.1	52.6	85.3	87.1	79.0
SFHPE (Peng, Zheng, and Chen 2023)	70.7	85.4	83.8	86.6	85.2	85.0	83.2	77.9	88.8	80.4	52.3	84.2	86.9	78.7
DA-LLPose (Ai et al. 2024)	70.9	85.7	84.1	86.9	85.5	85.3	83.5	78.2	89.1	80.7	52.6	84.5	87.2	79.0
AD-CLIP (Singha et al. 2023)	69.5	82.4	80.9	84.1	83.0	79.8	80.6	75.6	85.3	76.8	45.9	82.9	85.1	75.4
ReCLIP (Hu et al. 2024)	69.2	82.0	80.6	83.7	82.6	79.4	80.2	75.3	84.9	76.4	45.6	82.5	84.7	75.0
PDA (Phan et al. 2024)	70.4	86.3	83.6	86.1	84.7	85.5	82.7	78.4	88.3	80.9	51.8	84.7	86.4	79.2
CLIP2Pose (Ours)	**71.8**	**86.6**	**84.8**	**87.6**	**86.0**	**86.1**	**84.2**	**79.0**	**89.9**	**81.4**	**53.6**	**85.5**	**88.0**	**79.9**

Table 1: Comparison of our model with SOTA on human pose estimation task. The best results are highlighted in **bold**.

Method	RHD→WBH				
	MCP	PIP	DIP	Fin	All
Source only	15.2	20.1	20.3	22.4	19.7
CC-SSL	16.4	22.5	23.3	27.5	22.4
MDAM	19.0	25.2	26.0	30.2	25.1
RegDA	19.4	25.6	26.3	30.6	25.5
UniFrame	20.9	26.9	27.8	31.9	26.9
SFHPE	20.6	26.7	27.5	31.6	26.5
Ours	**25.2**	**31.2**	**32.1**	**36.2**	**31.2**

Table 2: Comparison of our model with SOTA on synthetic to wild setting. The best results are highlighted in **bold**.

Training Protocols

We train the entire model in an end-to-end fashion and the optimization follows a two-branch strategy: a supervised objective on the source domain and an adversarial alignment objective on the target domain. Supervised learning ensures accurate keypoint localization, while adversarial learning encourages domain-invariant feature representations.

Supervised Learning. Given source domain samples $x_s \in \mathcal{D}_s$ with annotated keypoint heatmaps H_{gt}, the predicted heatmap $\hat{H}_i$ is obtained by feeding the fused visual representation $\tilde{z}_i$ into the keypoint regression head. The regression loss is defined using the Mean Squared Error (MSE):

$$\mathcal{L}_{sup} = \frac{1}{N_s} \sum_{x_i^s \in \mathcal{D}_s} ||H_{gt} - H_i||^2. \tag{8}$$

where N_s denotes the total number of source samples.

Adversarial Learning. For unlabeled target samples $x_t \in \mathcal{D}_t$, we apply domain-adversarial training using a Gradient Reversal Layer (GRL) and a discriminator D_f that attempts to distinguish source from target features. The discriminator is trained with binary cross-entropy:

The discriminator is trained using the standard binary cross-entropy loss:

$$\mathcal{L}_D = -\mathbb{E}_{x_s}[\log D_f(\tilde{z}_s)] - \mathbb{E}_{x_t}[\log(1 - D_f(\tilde{z}_t))], \tag{9}$$

while the learnable parameters are updated adversarially via:

$$\mathcal{L}_{adv} = -\mathbb{E}_{x_t}[\log D_f(\tilde{z}_t)]. \tag{10}$$

Furthermore, we incorporate the directional alignment loss $\mathcal{L}_{align}$ introduced in Section to further constrain the semantic alignment. The final generator loss becomes:

$$\mathcal{L} = \mathcal{L}_{sup} + \mathcal{L}_{adv} + \lambda \cdot \mathcal{L}_{align}, \tag{11}$$

where λ balances the alignment constraint.

Experiments

Datasets and Evaluation Metrics

Datasets We evaluate our method on three human pose datasets and three hand pose datasets, encompassing both synthetic and real-world domains. For human pose estimation, we use SURREAL (Varol et al. 2017) as the synthetic source domain, while Human3.6M (Ionescu et al. 2013) and Leeds Sports Pose (LSP) (Johnson and Everingham 2010) serve as real-world target domains. For hand pose estimation, we adopt the Rendered Hand Pose Dataset (RHD) (Zimmermann and Brox 2017) as the synthetic source domain, with Hand-3D-Studio (H3D) (Zhao et al. 2020) and FreiHand (Zimmermann et al. 2019) as real-world target domains. More details are presented in Appendix.

Metrics We use the Percentage of Correct Keypoints (PCK) metric, a standard evaluation protocol in pose estimation. Specifically, we report PCK@0.05, which calculates the percentage of predicted keypoints that fall within 5% of the image size from the ground truth. For the 16-keypoint human body skeleton, we break down the results into six major regions: Shoulder (Sld), Elbow (Elb), Wrist, Hip, Knee, and Ankle. For the 21-keypoint hand skeleton, we report part-based accuracy on the Metacarpophalangeal joints (MCP), Proximal Interphalangeal joints (PIP), Distal Interphalangeal joints (DIP), and Fingertips (Fin).

Implementation Details We utilize a ResNet-101 backbone pre-trained on ImageNet as $\mathcal{F}_v$ and a transformer-based text encoder as $\mathcal{F}_t$. All parameters in $\mathcal{F}_v$ and $\mathcal{F}_t$ are frozen during training. We train only the SDP and SMVF modules using SGD with momentum 0.9 and weight decay $1e{-}4$, for 70 epochs. The batch size is set to 32, with 500 iterations per epoch. The learning rate starts at $1e{-}4$ and decays to $1e{-}5$ at epoch 45 and $1e{-}6$ at epoch 60. We adopt SimpleBaseline (Chen et al. 2020) as the keypoint estimation head. All

Method	RHD→H3D					RHD→FreiHand				
	MCP	PIP	DIP	Fin	All	MCP	PIP	DIP	Fin	All
Source only	67.7	64.5	63.6	55.0	62.1	35.4	50.4	55.1	51.0	47.1
CC-SSL (Mu et al. 2020)	81.5	79.9	74.4	64.0	75.1	37.4	48.2	50.1	46.5	43.8
MDAM (Li and Lee 2021)	82.3	79.6	72.3	61.5	74.1	32.3	48.1	51.7	47.3	45.1
RegDA (Jiang et al. 2021)	79.6	74.4	71.2	62.9	72.5	40.9	55.0	58.2	53.1	51.1
UniFrame (Kim et al. 2022)	86.7	84.6	78.9	68.1	79.6	43.5	64.0	67.4	62.4	58.5
SFHPE (Peng, Zheng, and Chen 2023)	88.4	89.2	80.9	71.4	82.2	43.7	65.9	66.6	63.1	58.8
DA-LLPose (Ai et al. 2024)	88.8	89.6	81.3	71.8	82.6	44.1	66.3	67.0	63.5	59.2
AD-CLIP (Singha et al. 2023)	86.8	86.9	76.4	69.7	79.9	40.9	63.3	63.9	61.2	56.5
ReCLIP (Hu et al. 2024)	88.9	89.8	81.2	71.7	82.9	43.7	65.2	66.6	63.1	58.9
PDA (Phan et al. 2024)	88.2	89.0	80.3	70.8	82.0	42.5	64.6	65.4	62.1	57.5
CLIP2Pose (Ours)	**89.8**	**90.3**	**82.0**	**72.3**	**83.6**	**44.8**	**66.4**	**68.0**	**64.4**	**59.9**

Table 3: Comparison of our method with SOTA on hand pose estimation task. The best results are highlighted in **bold**.

SDP	SMVF	DAL	Sld	Elb	Wrist	Hip	Knee	Ankle	All
✘	✘	✘	74.1	85.4	77.5	49.2	81.3	83.8	75.2
✔	✘	✘	76.1	87.2	79.3	51.4	83.0	85.7	77.1
✔	✔	✘	77.7	88.6	80.7	52.8	84.4	87.1	78.7
✔	✔	✔	**79.0**	**89.9**	**81.4**	**53.6**	**85.5**	**88.0**	**79.9**

Table 4: Ablation studies on SURREAL→Human3.6M

SDP	SMVF	DAL	MCP	PIP	DIP	Fin	All
✘	✘	✘	82.1	82.6	74.4	64.6	75.8
✔	✘	✘	85.1	85.9	77.7	66.3	78.9
✔	✔	✘	87.6	88.3	80.0	69.3	81.5
✔	✔	✔	**89.8**	**90.3**	**82.0**	**72.3**	**83.6**

Table 5: Ablation studies on RHD→H3D

results are averaged over five independent runs with different random seeds, and we report both the mean and standard deviation. Error bars in plots correspond to one standard deviation across the five runs. More details are in Appendix.

Comparisons with SOTA Methods

Human Body Pose Estimation We evaluate our method under two UDA settings: SURREAL→LSP and SURREAL→Human3.6M. As shown in Table 1, our method achieves the highest overall accuracy on both settings, reaching 83.6% on LSP and 79.9% on Human3.6M. Compared to traditional UDA methods, CLIP2Pose consistently improves performance with notable gains on challenging regions such as the wrist and ankle, which are typically harder to localize under domain shift. Compared to recent prompt-based UDA methods, which primarily target classification tasks, our approach is better suited for structured outputs like pose estimation. Existing methods often lack mechanisms to model fine-grained spatial dependencies, limiting their effectiveness in keypoint localization. In contrast, our structure aware prompts and semantic-guided visual interaction enable more accurate modeling of spatial hierarchies and domain-invariant features.

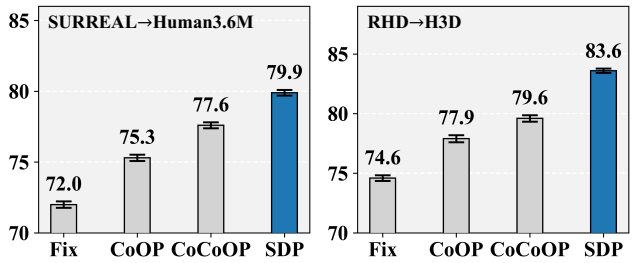

Figure 3: Comparison of different prompting methods.

Hand Pose Estimation We further validate our method on two hand pose estimation settings: RHD→H3D and RHD→FreiHand. As reported in Table 3, our model achieves 83.6% and 60.5% on H3D and FreiHand, respectively outperforming both conventional UDA and recent prompt-based methods. Beyond overall accuracy, our approach shows marked improvements on articulation-sensitive joints like PIP and DIP, which are particularly prone to occlusion. These results demonstrate that our structure aware prompting framework not only bridges the synthetic-to-real gap effectively, but also preserves spatial coherence essential for accurate pose estimation.

Synthetic to Wild Unlike human pose estimation, which benefits from in-the-wild datasets such as LSP, hand pose estimation still predominantly relies on lab-collected data. To address this limitation, we propose a more practical setting termed Synthetic to Wild. Specifically, we evaluate on RHD→WBH (Jin et al. 2020), where WBH comprises hand crops from the COCO dataset, featuring diverse appearances, complex backgrounds, and natural poses. As shown in Table 2, our method achieves the best performance. We also present additional results in a lab to wild setting and qualitative results on WBH in the Appendix.

Generalization to Unseen Domains In real-world applications, models often encounter domains that differ significantly from both source and target domains. To evaluate such scenarios, we assess domain generalization by testing models on entirely unseen datasets. First, we train on RHD→H3D and evaluate directly on FreiHAND. While

Method	FreiHand					Human3.6M						
	MCP	PIP	DIP	Fin	All	Sld	Elb	Wrist	Hip	Knee	Ankle	All
Source only	34.9	48.7	52.4	48.5	45.8	51.5	65.0	62.9	68.0	68.7	67.4	63.9
CCSSL	34.3	46.3	48.4	44.4	42.6	52.7	76.9	63.1	31.6	75.7	72.9	62.2
RegDA	37.8	51.8	53.2	47.5	46.9	76.9	80.2	69.7	52.0	80.3	80.0	73.2
UniFrame	35.6	52.3	55.4	50.6	47.1	77.0	85.9	73.8	47.6	80.7	80.6	74.3
CLIP2Pose (Ours)	**44.6**	**61.0**	**63.4**	**57.8**	**55.2**	**77.7**	**86.1**	**74.1**	**48.2**	**81.1**	**80.9**	**74.8**

Table 6: Domain generalization experiments on FreiHand and Human3.6M. The best results are highlighted in **bold**.

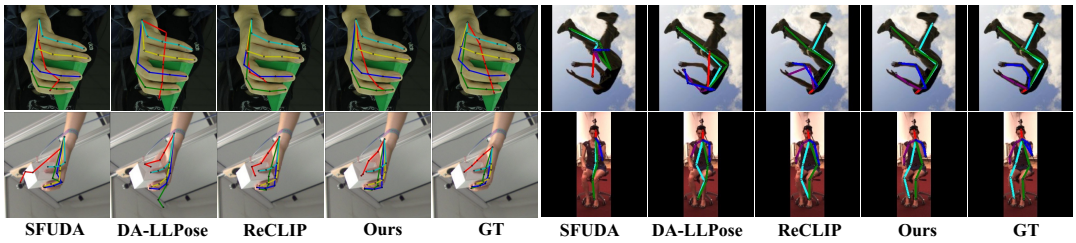

Figure 4: Qualitative results of our method. Top-left: H3D, Bottom-left: FreiHAND, Top-right: LSP, Bottom-right: Human3.6M.

most methods exhibit significant performance drops, our model achieves the best accuracy, particularly on PIP and DIP. Second, we train on SURREAL→LSP and test directly on Human3.6M. Despite LSP's limited size, our method reaches 74.8% accuracy, nearing adapted results. These results highlight the strong transferability of our structure aware design, even with minimal real-data supervision.

Qualitative Results We provide qualitative results in Figure 4. It is clear that our method produces more accurate keypoints compared to others. More qualitative results of synthetic to wild setting are available in Appendix.

Ablation Study

We conduct comprehensive ablation studies under two domain adaptation settings: SURREAL → Human3.6M for human pose estimation and RHD → H3D for hand pose estimation, as shown in Table 4 and Table 5, respectively. Our CLIP2Pose consists of three key components: Semantic-Driven Prompting (SDP), Semantic Modulation of Visual Features (SMVF), and the Directional Alignment Loss (DAL). Table 4 summarizes the impact of each component on the human pose estimation task. We observe steady improvements as each module is added. Starting from the baseline that includes only SDP, adding SMVF leads to an average improvement of 1.6 points. Incorporating DAL on top of both modules further increases the overall score to 79.9, resulting in a total gain of 4.8 points. The improvements are consistent across all metrics, with notable gains observed on the elbow, hip, and ankle, demonstrating the cumulative effect of each module. A similar trend is observed on the hand pose estimation task, as reported in Table 5. The addition of SMVF improves the average accuracy from 78.9 to 81.5. When DAL is also included, the performance reaches 83.6, yielding a total gain of 7.8 points over the baseline. More discussion is provided in the Appendix.

Figure 3 compares different prompting strategies under the same setting. Our proposed SDP consistently exceeds CoOP and CoCoOP in Human3.6M and H3D. These results demonstrate the effectiveness of task-aware prompting for cross-domain generalization.

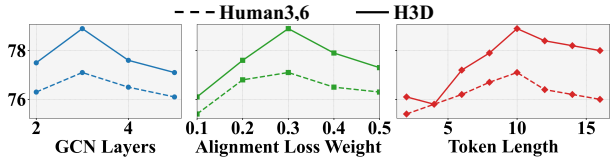

Figure 5: Parameter Analysis on RHD→H3D and SURREAL→Human3.6M.

Parameter Analysis We analyze the impact of GCN layers n, alignment loss weight λ, and token length m, as illustrated in Figure 5. Performance peaks at $n = 3$, $\lambda = 0.3$, and $m = 10$, showing consistent trends across both H3D and Human3.6M. Deviations from these values result in only minor performance drops, indicating that the method is robust to changes in hyperparameters.

Conclusion

We proposed CLIP2Pose, a framework for unsupervised domain adaptive pose estimation that leverages a frozen CLIP model to guide the learning of domain-invariant representations. By incorporating structural, stylistic, and instance-specific cues into semantically informed prompts, CLIP2Pose exploits the representational power of CLIP to align visual features with high-level structural semantics. This alignment steers the feature space toward domain-agnostic and task relevant dimensions, resulting in more robust keypoint localization across domains. Extensive experiments on both hand and body pose benchmarks demonstrate that CLIP2Pose consistently outperforms existing domain adaptation and prompt-based methods.

Acknowledgments

This paper is supported by National Natural Foundation of China (No. 62207014).

References

Ai, Y.; Qi, Y.; Wang, B.; Cheng, Y.; Wang, X.; and Tan, R. T. 2024. Domain-adaptive 2d human pose estimation via dual teachers in extremely low-light conditions. In *European Conference on Computer Vision*, 221–239. Springer.

Bai, S.; Zhang, M.; Zhou, W.; Huang, S.; Luan, Z.; Wang, D.; and Chen, B. 2024a. Prompt-based distribution alignment for unsupervised domain adaptation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 729–737.

Bai, S.; Zhou, W.; Luan, Z.; Wang, D.; and Chen, B. 2024b. Improving cross-domain few-shot classification with multilayer perceptron. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 5250–5254. IEEE.

Cao, Z.; Radosavovic, I.; Kanazawa, A.; and Malik, J. 2021. Reconstructing hand-object interactions in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 12417–12426.

Chen, T.; Kornblith, S.; Norouzi, M.; and Hinton, G. 2020. A simple framework for contrastive learning of visual representations. In *International Conference on Machine Learning*, 1597–1607.

Chen, Z.; Ge, J.; Zhan, H.; Huang, S.; and Wang, D. 2021. Pareto self-supervised training for few-shot learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 13663–13672.

Deng, Q.; Liu, Y.; Li, W.; and Wang, G. 2024. Structure-Aware Human Body Reshaping with Adaptive Affinity-Graph Network. *ArXiv preprint arXiv:2404.13983*.

Doosti, B.; Naha, S.; Mirbagheri, M.; and Crandall, D. J. 2020. Hope-net: A graph-based model for hand-object pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6608–6617.

Gao, P.; Geng, S.; Zhang, R.; Ma, T.; Fang, R.; Zhang, Y.; Li, H.; and Qiao, Y. 2024. Clip-adapter: Better vision-language models with feature adapters. *International Journal of Computer Vision*, 132(2): 581–595.

Ge, C.; Huang, R.; Xie, M.; Lai, Z.; Song, S.; Li, S.; and Huang, G. 2023. Domain adaptation via prompt learning. *IEEE Transactions on Neural Networks and Learning Systems*.

Guo, Z.; Zhang, R.; Qiu, L.; Ma, X.; Miao, X.; He, X.; and Cui, B. 2023. Calip: Zero-shot enhancement of clip with parameter-free attention. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, 746–754.

Han, Z.; Sun, H.; and Yin, Y. 2022. Learning transferable parameters for unsupervised domain adaptation. *IEEE Transactions on Image Processing*, 31: 6424–6439.

Hu, X.; Zhang, K.; Xia, L.; Chen, A.; Luo, J.; Sun, Y.; Wang, K.; Qiao, N.; Zeng, X.; Sun, M.; et al. 2024. Reclip: Refine contrastive language image pre-training with source free domain adaptation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2994–3003.

Huang, Z.; Zeng, Z.; Huang, Y.; Liu, B.; Fu, D.; and Fu, J. 2021. Seeing out of the box: End-to-end pre-training for vision-language representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 12976–12985.

Ionescu, C.; Papava, D.; Olaru, V.; and Sminchisescu, C. 2013. 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.

Jia, C.; Yang, Y.; Xia, Y.; Chen, Y.-T.; Parekh, Z.; Pham, H.; Le, Q.; Sung, Y.-H.; Li, Z.; and Duerig, T. 2021. Scaling up visual and vision-language representation learning with noisy text supervision. In *International Conference on Machine Learning*, 4904–4916. PMLR.

Jiang, J.; Ji, Y.; Wang, X.; Liu, Y.; Wang, J.; and Long, M. 2021. Regressive domain adaptation for unsupervised keypoint detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6780–6789.

Jin, R.; Zhang, J.; Yang, J.; and Tao, D. 2022. Multibranch adversarial regression for domain adaptative hand pose estimation. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(9): 6125–6136.

Jin, S.; Xu, L.; Xu, J.; Wang, C.; Liu, W.; Qian, C.; Ouyang, W.; and Luo, P. 2020. Whole-body human pose estimation in the wild. In *Proceedings of the European Conference on Computer Vision*, 196–214.

Johnson, S.; and Everingham, M. 2010. Clustered pose and nonlinear appearance models for human pose estimation. In *BMCV*, volume 2, 5. Aberystwyth, UK.

Khattak, M. U.; Rasheed, H.; Maaz, M.; Khan, S.; and Khan, F. S. 2023. Maple: Multi-modal prompt learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 19113–19122.

Kim, D.; Wang, K.; Saenko, K.; Betke, M.; and Sclaroff, S. 2022. A unified framework for domain adaptive pose estimation. In *Proceedings of the European Conference on Computer Vision*, 603–620.

Kim, W.; Son, B.; and Kim, I. 2021. Vilt: Vision-and-language transformer without convolution or region supervision. In *International conference on machine learning*, 5583–5594. PMLR.

Kumarlal Fernando, C. 1975. The Physiology of the Joints Vol 3, The Trunk and the Vertebral Column.

Lai, Z.; Vesdapunt, N.; Zhou, N.; Wu, J.; Huynh, C. P.; Li, X.; Fu, K. K.; and Chuah, C.-N. 2023. Padclip: Pseudo-labeling with adaptive debiasing in clip for unsupervised domain adaptation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 16155–16165.

Li, C.; and Lee, G. H. 2021. From synthetic to real: Unsupervised domain adaptation for animal pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1482–1491.

Li, L.; Tian, L.; Zhang, X.; Wang, Q.; Zhang, B.; Bo, L.; Liu, M.; and Chen, C. 2023. Renderih: A large-scale synthetic dataset for 3d interacting hand pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 20395–20405.

Lin, Q.; Yang, L.; and Yao, A. 2023. Cross-domain 3d hand pose estimation with dual modalities. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 17184–17193.

Mu, J.; Qiu, W.; Hager, G. D.; and Yuille, A. L. 2020. Learning from synthetic animals. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 12386–12395.

Ohkawa, T.; Li, Y.-J.; Fu, Q.; Furuta, R.; Kitani, K. M.; and Sato, Y. 2022. Domain adaptive hand keypoint and pixel localization in the wild. In *Proceedings of the European Conference on Computer Vision*, 68–87.

Peng, Q.; Zheng, C.; and Chen, C. 2023. Source-free domain adaptive human pose estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 4826–4836.

Phan, V. H.; Tran, T. L.; Tran, Q.; and Le, T. 2024. Enhancing domain adaptation through prompt gradient alignment. *Advances in Neural Information Processing Systems*, 37: 45518–45551.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763. PmLR.

Singha, M.; Pal, H.; Jha, A.; and Banerjee, B. 2023. Ad-clip: Adapting domains in prompt space using clip. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 4355–4364.

Varol, G.; Romero, J.; Martin, X.; Mahmood, N.; Black, M. J.; Laptev, I.; and Schmid, C. 2017. Learning from synthetic humans. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 109–117.

Wu, Y.; Xing, M.; Zhang, Y.; Xie, Y.; and Qu, Y. 2024. Clip2uda: Making frozen clip reward unsupervised domain adaptation in 3d semantic segmentation. In *Proceedings of the ACM International Conference on Multimedia*, 8662–8671.

Yuan, S.; Garcia-Hernando, G.; Stenger, B.; Moon, G.; Chang, J. Y.; Lee, K. M.; Molchanov, P.; Kautz, J.; Honari, S.; Ge, L.; et al. 2018. Depth-based 3d hand pose estimation: From current achievements to future goals. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2636–2645.

Zhang, M.; Huang, S.; Li, W.; and Wang, D. 2022a. Tree structure-aware few-shot image classification via hierarchical aggregation. In *European Conference on Computer Vision*, 453–470. Springer.

Zhang, M.; Yuan, J.; He, Y.; Li, W.; Chen, Z.; and Kuang, K. 2023. Map: Towards balanced generalization of iid and ood through model-agnostic adapters. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 11921–11931.

Zhang, P.; Zhang, B.; Chen, D.; Yuan, L.; and Wen, F. 2020. Cross-domain correspondence learning for exemplar-based image translation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 5143–5153.

Zhang, R.; Zhang, W.; Fang, R.; Gao, P.; Li, K.; Dai, J.; Qiao, Y.; and Li, H. 2022b. Tip-adapter: Training-free adaption of clip for few-shot classification. In *European Conference on Computer Vision*, 493–510. Springer.

Zhao, M.; Zhu, Z.; Shi, J.; Wang, Z.; Chen, J.; An, H.; and Yan, B. 2025. PromptSeg: Learning to Segment Medical Image via Visual Prompts. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, 1–5. IEEE.

Zhao, Z.; Wang, T.; Xia, S.; and Wang, Y. 2020. Hand-3d-Studio: A New Multi-View System for 3d Hand Reconstruction. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2478–2482.

Zhou, K.; Yang, J.; Loy, C. C.; and Liu, Z. 2022. Cocoop: Conditional prompt learning for vision-language models. *ArXiv preprint arXiv:2203.05557*.

Zhu, B.; Niu, Y.; Han, Y.; Wu, Y.; and Zhang, H. 2023. Prompt-aligned gradient for prompt tuning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 15659–15669.

Zhu, J.; Chen, Y.; and Wang, L. 2024. CLIP the Divergence: Language-guided Unsupervised Domain Adaptation. *arXiv preprint arXiv:2407.01842*.

Zimmermann, C.; and Brox, T. 2017. Learning to estimate 3d hand pose from single rgb images. In *Proceedings of the IEEE International Conference on Computer Vision*, 4903–4911.

Zimmermann, C.; Ceylan, D.; Yang, J.; Russell, B.; Argus, M.; and Brox, T. 2019. Freihand: A dataset for markerless capture of hand pose and shape from single rgb images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 813–822.