# Vision-aware Multimodal Prompt Tuning for Uploadable Multi-source Few-shot Domain Adaptation

**Kuanghong Liu, Jin Wang*, Kangjian He*, Dan Xu, Xuejie Zhang**

School of Information Science and Engineering, Yunnan University, Kunming, China
liukh@mail.ynu.edu.cn, {wangjin, hekj, danxu, xjzhang}@ynu.edu.cn

## Abstract

Conventional multi-source domain few-shot adaptation (MFDA) faces the challenge of further reducing the load on edge-side devices in low-resource scenarios. Considering the native language-supervised advantage of CLIP and the plug-and-play nature of prompt to transfer CLIP efficiently, this paper introduces an uploadable multi-source few-shot domain adaptation (UMFDA) schema. It belongs to a decentralized edge collaborative learning in the edge-side models that must maintain a low computational load. And only a limited amount of annotations in source domain data is provided, with most of the data being unannotated. Further, this paper proposes a vision-aware multimodal prompt tuning framework (VAMP) under the decentralized schema, where the vision-aware prompt guides the text domain-specific prompt to maintain semantic discriminability and perceive the domain information. The cross-modal semantic and domain distribution alignment losses optimize each edge-side model, while text classifier consistency and semantic diversity losses promote collaborative learning among edge-side models. Extensive experiments were conducted on OfficeHome and DomainNet datasets to demonstrate the effectiveness of the proposed VAMP in the UMFDA, which outperformed the previous prompt tuning methods.

**Code** — https://github.com/lkh-meredith/VAMP-UMFDA

## Introduction

Multi-source few-shot domain adaptation (MFDA) is a resource-limited multi-source domain adaptation scenario, since large-scale manual annotations of each source domain are laborious and difficult, especially for the disease dataset that needs expert labeling or is even inaccessible when involves private data (Gulshan et al. 2016; Harmon et al. 2020). That means only a limited amount of annotations in the source domain data is provided, with most of the data being unannotated. Therefore, it is more practical and worth further exploration (Yue et al. 2021a; Kim et al. 2020). However, the improvement of the conventional MFDA method comes at the expense of storing additional clustering prototype features as a classifier (Yue et al. 2021a). In practice, this might increase the cost of the storage and computation for edge devices (McMahan et al. 2016).

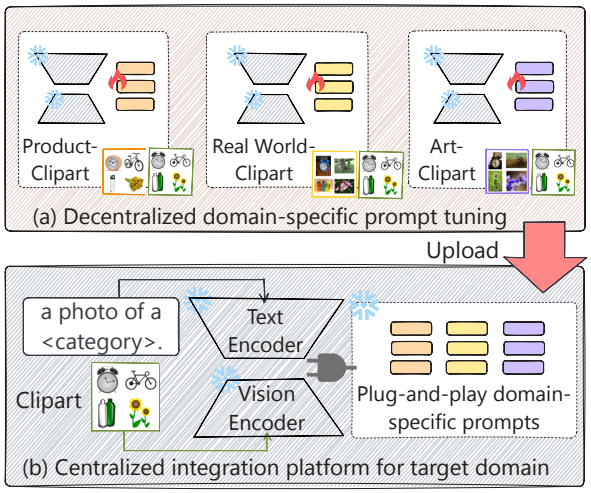



Figure 1: The illustration of uploadable multi-source few-shot domain adaptation (UMFDA) schema for decentralized edge learning.

The pretrained vision-and-language model (VLM), such as CLIP (Radford et al. 2021), has attracted attention for its remarkable zero-shot inference performance and transferability. Pretraining by cross-modal alignment of contrastive learning, CLIP has the native advantage of language-guided supervision to form similar semantic clusters. A critical insight is to leverage a manual-craft text prompt, e.g., `a photo of a <category>.`, as a query prompt for the text encoder to inspire CLIP's potential. With the development of prompt tuning in NLP (Lester, Al-Rfou, and Constant 2021; Li and Liang 2021; Liu et al. 2021; Shin et al. 2020), some methods (Zhou et al. 2022a; Zang et al. 2022; Chen et al. 2022; Lu et al. 2022; Khattak et al. 2023) replaced the manual-crafted prompt with a small fraction of learnable parameters. Only by updating the fewer prompt parameters rather than the entire model will these plug-and-play learnable prompts make CLIP favorably adaptable to various downstream tasks, even under the few annotated sample conditions. With insights into the native language-supervised benefit of CLIP and the prompt's efficient plug-

*Corresponding author.

and-play capability, they are well-suited for application in low-resource, low-load edge computing scenarios. Therefore, we introduce an uploadable multi-source few-shot domain adaptation (UMFDA) schema for decentralized edge learning in this paper, as illustrated in Figure 1.

Specifically, Figure 1(a) is the decentralized training stage for edge-side devices. By domain-specific prompt tuning, these edge-side devices perform lightweight edge collaborative learning. It avoids the need to finetune the entire domain-specific model and reduces the burden of storing and processing trainable parameters (Zhao et al. 2024). Figure 1(b) is the centralized integration platform that accepts all the uploaded domain-specific prompts. Once these trained domain-specific prompts are inserted into the frozen CLIP, the domain-specific image extractor and text classifier are constructed, where the image encoder of CLIP serves as a feature extractor; the text encoder of CLIP is regarded as a text classifier. The target domain's results are integrated from directly inferences of all the domain-specific models without further training in the centralized platform.

Nevertheless, the current prompt tuning technologies are not suitable for the UMFDA. As summarized in Figure 2(a), the prevalent prompt tuning methods (Zhou et al. 2022a; Jia et al. 2022; Khattak et al. 2023) are domain-agnostic and neglect domain shifts and distribution differences among domains. Figure 2(b) represents another recently emerged domain adaption prompt tuning method (Ge et al. 2022; Chen et al. 2023). They disentangle context prompts as domain-agnostic and domain-specific prompts to embed domain information into text prompts by contrasting the source and target domain data pair. However, this approach increases the central equipment's computing load, as additional training is still required to centrally derive a domain-invariant shared representation space among the learned individual prompts (Chen et al. 2023). Additionally, they only affect the text encoder when adapting to changing domains, while learning domain-specific discriminant image features is difficult. It may damage the distribution of the representation in CLIP and cause a loss of the learned semantic information (Singha et al. 2023; Bai et al. 2024; Du et al. 2024).

This study proposes a vision-aware multimodal prompt tuning framework (VAMP) for the UMFDA, as shown in Figure 2(c). The multimodal prompt tuning is applied to maintain the discriminative ability of the learned features. Notably, the difference from the previous prompt-agnostic multimodal prompt (*text* $\rightarrow$ *vision*, Figure 2(a)) is that the visual prompt perceives the domain information and is then projected to the text prompt used for the specific domain (*vision* $\rightarrow$ *text*). By doing so, these domain-specific text prompts guided by visual perception can be searched to describe the domain images, rather than being the manual designing or using vision-independent text prompts. The original manual text prompt, i.e., `A photo of a <category>.`, as the domain-agnostic prompt is concatenated with the domain-specific text prompt to preserve the generalization knowledge of CLIP.

To optimize this framework in a decentralized manner, the pairs source and target domain data are fully utilized. Each edge-side model with domain-specific vision-aware multimodal prompts is trained internally, while engaging in collaborative learning among the edge-side models. Concretely, cross-modal semantic alignment (CSA) and domain distribution alignment (DDA) losses are used in each edge-side model; text classifier consistency (TCC) and text semantic diversity (TSD) losses are introduced to facilitate collaborative learning among multiple edge-side models. The main contribution can be summarized as follows:

- Inspired by the plug-and-play prompts to transfer the CLIP efficiently, this study introduces a UMFDA schema for low-resource and low-load edge learning and further proposes the VAMP framework. The customized domain-specific prompts in the VAMP are vision-aware multimodal prompts, where vision-aware prompts guide domain-specific text prompts.
- VAMP is optimized in a decentralized training manner by four different losses. CSA and DDA losses ensure cross-modal semantic information and distribution alignments within edge-side models; TCC and TSD losses facilitate collaborative learning among edge-side models.
- Extensive experiments on OfficeHome and DomainNet datasets demonstrate the effectiveness of the VAMP, which outperforms the previous prompt tuning methods.

## Preliminaries

**MFDA.** The UMFDA also follows the setting of MFDA from MFSAN (Yue et al. 2021a), which focuses on transferring generalization knowledge from the multiple source domains to the target domain. In the scenario, there is a small annotated data $D^i_{s\_a} = \{(x^{a,i}_j, y^{a,i}_j)\}^{N^i_a}_{j=1}$ and a large unannotated data $D^i_{s\_u} = \{x^{u,i}_j\}^{N^i_u}_{j=1}$ for the $i$-th source domain $D^i_s = D^i_{s\_a} \cup D^i_{s\_u}, i \in \{1, 2, ..., M\}$, where $N^i_a$ and $N^i_u$ ($N^i_a \ll N^i_u$) are the size of the annotated and unannotated samples, respectively. $M$ is the number of source domains. $D_t = \{x^t_j\}^{N_t}_{j=1}$ denotes the target dataset without label annotation, where $N_t$ represents the count of target samples. Notably, it assumes that the data of different domains come from different distributions and all share the same label space. The objective is to train a domain adaptation model on multiple domain data $D_s$ and $D_t$, which enables prediction labels on the target samples as correctly as possible.

**CLIP Inference.** Benefiting from the alignment pretraining of language and vision modalities on large-scale text-image pairs by contrastive learning, the CLIP model has achieved outstanding zero-shot inference. A piece of manual-crafted text prompts, i.e., `a photo of a <category>.`, is enough to inspire CLIP's potential (Radford et al. 2021). Given an image $x$, its visual representation $z \in \mathbb{R}^d$ and text representations $W = \{w_1, w_2, ..., w_K\} \in \mathbb{R}^{K \times d}$ for the $K$ candidate categories are produced by the CLIP's vision encoder $\Psi$ and text encoder $\Phi$, respectively. The probability that the image belongs to the $c$-th class is calculated as,

$$p(\hat{y} = c|x) = \frac{\exp(\cos(w_c, z)/\mathrm{T})}{\sum_{k=1}^{K} \exp(\cos(w_k, z)/\mathrm{T})} \quad (1)$$

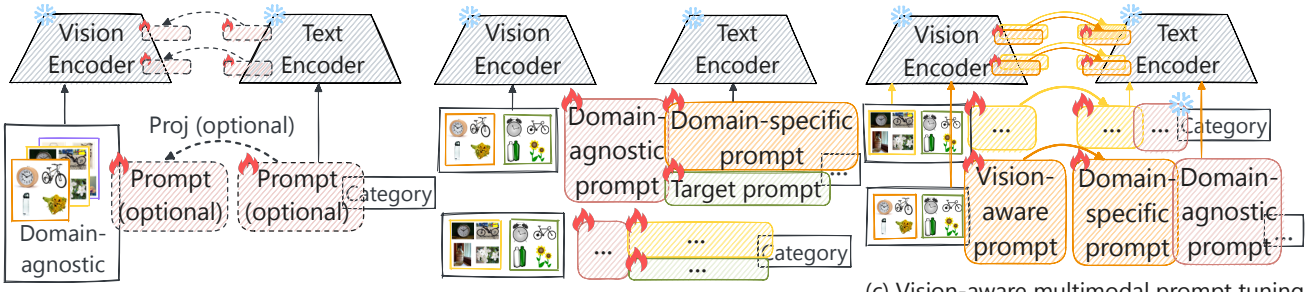



Figure 2: Summary of various prompt tuning technologies (best viewed in color). (a) concludes the several prevalent prompt tuning methods while they are domain-agnostic. (b) represents the typical prompt tuning methods of single-source domain adaptation focusing on disentangling the prompts to explore the difference between the source and target domains. It must be further aligned in the center device among multiple source domains. (c) is our proposed vision-aware multimodal prompt tuning method tailored for the UMFDA.

where $\cos(\cdot,\cdot)$ denotes the cosine similarity and T is a fixed temperature parameter learned by CLIP.

**Multimodal Prompt Tuning.** MaPLe (Khattak et al. 2023) introduced a coupling function to enhance interaction with the prompts of vision and text modality for synergic optimization in the CLIP. Specifically, a series of new learnable prompts $\mathbf{p} = [\mathbf{p}_1, \mathbf{p}_2, ..., \mathbf{p}_b] \in \mathbb{R}^{b\times d_T}$ are introduced in each transformer layer of $\Phi$, up to the depth $J$. Vision prompts $\tilde{\mathbf{p}}^l$ are obtained by projecting $\mathbf{p}^l$ via a coupling function $\text{proj}^l(\cdot)$ at $l$-th transformer layer, i.e., $\tilde{\mathbf{p}}^l = \text{proj}^l(\mathbf{p}^l)$. The $\text{proj}^l(\cdot)$ is implemented as one linear layer that maps dimension $d_T$ to $d_V$. Supposing that both encoder $\Phi$ and $\Psi$ have $L$ transformer layers and image patch embeddings are $\mathbf{q^0} = [\mathbf{q}_1^0, \mathbf{q}_2^0, ..., \mathbf{q}_s^0] \in \mathbb{R}^{s\times d_V}$, the entire process can be written as,

$$\begin{aligned}
[\mathbf{c}_T^l, _, \mathbf{e}^l] &= f_\Phi^l([\mathbf{c}_T^{l-1}, \mathbf{p}^{l-1}, \mathbf{e}^{l-1}]),\ l = 1, 2, \ldots, J \\
[\mathbf{c}_T^m, \mathbf{p}^m, \mathbf{e}^m] &= f_\Phi^m([\mathbf{c}_T^{m-1}, \mathbf{p}^{m-1}, \mathbf{e}^{m-1}]),\ m = J+1, \ldots, L \\
w &= \text{Proj}_\Phi(e_n^L)
\end{aligned} \tag{2}$$

$$\begin{aligned}
[\mathbf{c}_V^l, \mathbf{q}^l, _] &= f_\Psi^l([\mathbf{c}_V^{l-1}, \mathbf{q}^{l-1}, \text{proj}^{l-1}(\mathbf{p}^{l-1})]), l = 1, 2, ..., J \\
[\mathbf{c}_V^m, \mathbf{q}^m, \tilde{\mathbf{p}}^m] &= f_\Psi^m([\mathbf{c}_V^{m-1}, \mathbf{q}^{m-1}, \tilde{\mathbf{p}}^{m-1}]), m = J+1, ..., L \\
z &= \text{Proj}_\Psi(\mathbf{c}_V^L)
\end{aligned} \tag{3}$$

where $\mathbf{c}_T \in \mathbb{R}^{d_T}$ and $\mathbf{c}_V \in \mathbb{R}^{d_V}$ are the starting token embeddings. $\mathbf{e^0} = [e_1^0, e_2^0, ..., e_n^0] \in \mathbb{R}^{n\times d_T}$ are the fixed token embeddings including category label. $[\cdot,\cdot]$ represents the concatenation operation. Each text representation $w \in \mathbb{R}^d$ and image representation $z \in \mathbb{R}^d$ are finally projected to a common vision-and-language space via $\text{Proj}_\Phi(\cdot)$ and $\text{Proj}_\Psi(\cdot)$, respectively.

# Method

## Vision-aware Multimodal Prompt Tuning

In the UMFDA setting, we still adopt multimodal prompts as domain-specific prompts to maintain semantic discriminability. Differently, the domain information is perceived by visual prompts during tuning the image extractor. Then, these visual prompts are projected as learnable domain-specific text prompts concatenated with the manual-crafted prompt to search optimal text queries in the corresponding domain. Therefore, the reverse of the coupling function in MaPLe, i.e.,$\mathbf{p}^l = \text{proj}^l(\tilde{\mathbf{p}}^l)$, our vision-aware prompt tuning is written as,

$$\begin{aligned}
[\mathbf{c}_V^l, \mathbf{q}^l, _] &= f_\Psi^l([\mathbf{c}_V^{l-1}, \mathbf{q}^{l-1}, \tilde{\mathbf{p}}^{l-1}]), l = 1, 2, ..., J \\
[\mathbf{c}_V^m, \mathbf{q}^m, \tilde{\mathbf{p}}^m] &= f_\Psi^m([\mathbf{c}_V^{m-1}, \mathbf{q}^{m-1}, \tilde{\mathbf{p}}^{m-1}]), m = J+1, ..., L \\
z &= \text{Proj}_\Psi(\mathbf{c}_V^L)
\end{aligned} \tag{4}$$

$$\begin{aligned}
[\mathbf{c}_T^l, _, \mathbf{e}^l] &= f_\Phi^l([\mathbf{c}_T^{l-1}, \text{proj}^{l-1}(\tilde{\mathbf{p}}^{l-1}), \mathbf{e}^{l-1}]), l = 1, 2, ..., J \\
[\mathbf{c}_T^m, \mathbf{p}^m, \mathbf{e}^m] &= f_\Phi^m([\mathbf{c}_T^{m-1}, \mathbf{p}^{m-1}, \mathbf{e}^{m-1}]), m = J+1, ..., L \\
w &= \text{Proj}_\Phi(e_n^L)
\end{aligned} \tag{5}$$

where $\mathbf{e}^0$ denotes the fixed embeddings of the original text prompt, i.e., "`a photo of a <category>.`", which is seen as the domain-agnostic prompt.

## VAMP Framework

Figure 3 is the conceptual diagram of the proposed VAMP framework, which aims to learn a group of trainable domain-specific vision-aware multimodal prompts $\{\mathbf{P}_1, \mathbf{P}_2, ..., \mathbf{P}_M\}$ (including newly inserted learnable prompts for each layer and their corresponding projection functions) for different edge-side models in decentralized training way. Domain-specific image extractor and text classifier for source domain $i$ are thus represented as $\Psi_i(\cdot, \mathbf{P}_i)$ and $\Phi_i(\mathbf{E}, \mathbf{P}_i)$, respectively, where $\mathbf{E} = \{\mathbf{e}_1, \mathbf{e}_2, ..., \mathbf{e}_K\} \in \mathbb{R}^{K\times n\times d_T}$ denotes the fixed token embeddings of all categories. In this framework, $M$ domain-alignment vision representation spaces are learned by $\{\Psi_i(\cdot, \mathbf{P}_i)\}_{i=1}^M$, rather than learning a common domain-invariant feature space by aligning multiple domains in the subsequent centralized integration stage, as illustrated in Figure 3 (a). The overall decentralized training framework VAMP is shown in Figure 3 (b), where the CSA and DDA losses are used in each edge-side model; the TCC and TSD losses are introduced to facili-

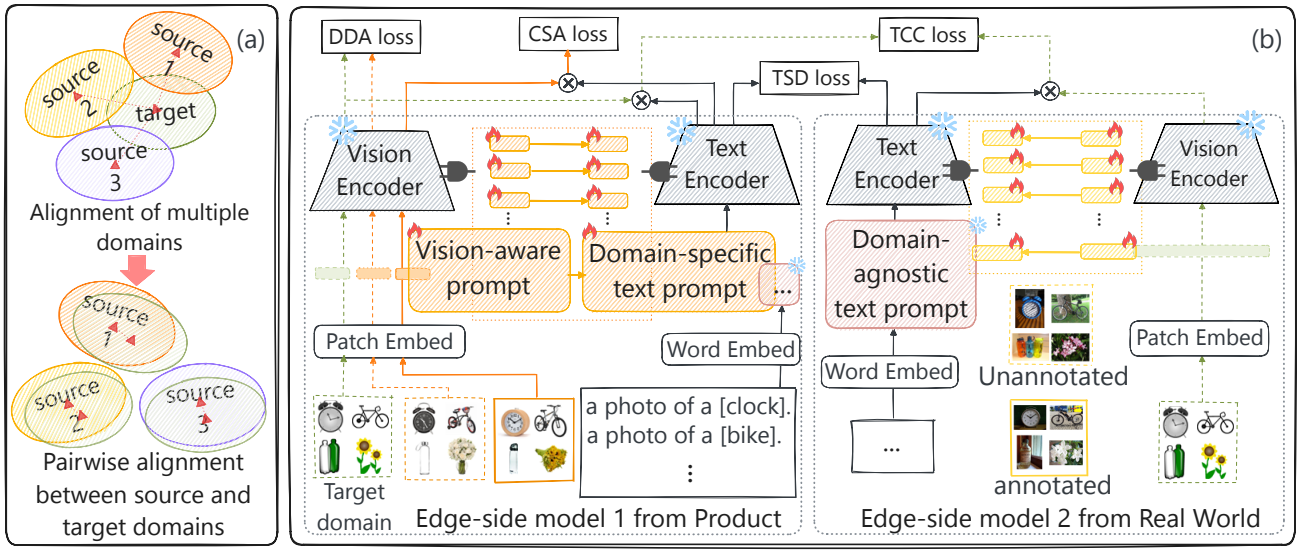



Figure 3: (a) Illustration of a conceptual diagram of domain alignment. The first alignment approach at the top, multiple source domains aligning with the target domain, is unsuitable for the decentralized edge computing scenario because the centralized model needs more aligning training. It is also hard to match all distributions of source and target domains. The second idea inspired by Zhu et al. (2019), pair-wise alignment between source and target domains, is thus adopted by the VAMP framework. (b) The proposed decentralized training framework VAMP. For clarity, only two edge-side models are drawn here.

tate collaborative learning among multiple edge-side models. They are described in detail as follows.

**Cross-modal Semantic Alignment.** For each domain-specific model, a batch of annotated data $\{(x_j^{a,i}, y_j^{a,i})\}_{j=1}^{n_a^i} \subset D_{s\_a}^i$ and unannotated target data $\{x_j^t\}_{j=1}^{n_t} \subset D_t$ are sampled and applied to keep the semantic discriminability by cross-modal contrastive learning. For the annotated source samples, the model can be directly optimized with ground-truth labels as,

$$\mathcal{L}_{CSA}^a = -\frac{1}{n_a^i}\sum_{j=1}^{n_a^i}\sum_{k=1}^{K}\left[y_{j,k}^{a,i}\cdot\log p(\hat{y}_{j,k}^{a,i}|x_j^{a,i};\mathbf{P}_i)\right] \tag{6}$$

where $y_j^{a,i}$ is the one-hot ground-truth label. $p(\hat{y}=c|x;\mathbf{P}_i)$ of an image $x$ categorizing to the $c$-th class is rewritten as,

$$p(\hat{y}=c|x;\mathbf{P}_i) = \frac{\exp(\cos(\Phi(\mathbf{E}_c,\mathbf{P}_i),\Psi(x,\mathbf{P}_i))/\mathrm{T})}{\sum_{k=1}^{K}\exp(\cos(\Phi(\mathbf{E}_k,\mathbf{P}_i),\Psi(x,\mathbf{P}_i))/\mathrm{T})} \tag{7}$$

Meanwhile, by zero-shot inference of CLIP as shown in Eq.(1), a qualified pseudo label of target sample whose maximum prediction probability $\hat{\tau}$ is larger than a fixed threshold $\mathcal{T}$ is generated. When the number of qualified pseudo labels is not equal to 0 in a batch,

$$\mathcal{L}_{CSA}^t = -\frac{\sum_{j=1}^{n_t}\mathbb{I}\{\hat{\tau}_j\geq\mathcal{T}\}\sum_{k=1}^{K}\left[\hat{y}_{j,k}^{\mathrm{zs}}\cdot\log p_i(\hat{y}_{j,k}^t|x_j^t;\mathbf{P}_i)\right]}{\sum_{j=1}^{n_t}\mathbb{I}\{\hat{\tau}_j\geq\mathcal{T}\}} \tag{8}$$

where $\mathbb{I}\{\cdot\}$ is an indicator function. $\hat{y}_j^{\mathrm{zs}}$ is a one-hot pseudo label predicted by zero-shot CLIP. Therefore, the cross-modal semantic alignment loss is written as,

$$\mathcal{L}_{CSA} = \mathcal{L}_{CSA}^a + \mathcal{L}_{CSA}^t \tag{9}$$

**Domain Distribution Alignment.** To align the distribution for each pair of source and target domain, maximum mean discrepancy (MMD) (Gretton et al. 2007, 2012) is used to optimize by exploiting extensive unannotated data from $D_{s\_u}^i$ and $D_t$. Supposed that the image features $\{z_j^{u,i}\}_{j=1}^{N_u^i} = \{\Psi_i(x_j^{u,i},\mathbf{P}_i)\}_{j=1}^{N_u^i}$ and $\{z_j^t\}_{j=1}^{N_t} = \{\Psi_i(x_j^t,\mathbf{P}_i)\}_{j=1}^{N_t}$ are independently and identically drawn from $P_s^i$ and $Q_t$, respectively. MMD works by distinguishing statistical hypothesis testing of two samples that if they are similar then they are likely from the same distribution. To measure the difference between $P_s^i$ and $Q_t$, the squared MMD with kernel mean embeddings is formulated as,

$$\mathrm{MMD}^2(\mathcal{F},P_s^i,Q_t) \triangleq \left\|\mathbb{E}_{z^{u,i}\sim P_s^i}\left[\phi(z^{u,i})\right]-\mathbb{E}_{z^t\sim Q_t}\left[\phi(z^t)\right]\right\|_{\mathcal{H}}^2 \tag{10}$$

where the function class $\mathcal{F}$ is the unit ball in a reproducing kernel Hilbert space (RKHS) $\mathcal{H}$ endowed with a characteristic kernel $\kappa$. $\phi(\cdot)$ is feature mapping that maps into $\mathcal{H}$ and $\kappa$ denotes $\kappa(z^{u,i},z^t) = \left\langle\phi(z^{u,i}),\phi(z^t)\right\rangle_{\mathcal{H}}$, where $\langle\cdot,\cdot\rangle$ is the inner product of vectors. Practically, given a batch $n_u^i$ of source unannotated samples and $n_t$ target samples, the form of empirical estimates with finite samples is calculated as,

$$\begin{aligned}\widehat{\mathrm{MMD}}^2(\mathcal{F},P_s^i,Q_t) &= \frac{1}{n_u^i\cdot n_u^i}\sum_{j=1}^{n_u^i}\sum_{k\neq j}^{n_u^i}\kappa(z_j^{u,i},z_k^{u,i})\\ &+\frac{1}{n_t\cdot n_t}\sum_{j=1}^{n_t}\sum_{k\neq j}^{n_t}\kappa(z_j^t,z_k^t)-\frac{2}{n_u^i\cdot n_t}\sum_{j=1}^{n_u^i}\sum_{k=1}^{n_t}\kappa(z_j^{u,i},z_k^t)\end{aligned} \tag{11}$$

where $\kappa$ is the universal kernel such as Gaussian kernel $\kappa(z,z') = \exp\left(-\frac{1}{2\sigma}|z-z'|^2\right)$ with bandwidth $\sigma$. When

the feature space is a universal RKHS, MMD is 0 if and only if $P_s^i = Q_t$. Therefore, minimizing MMD under this condition is equivalent to minimizing the distance between all moments of the two distributions $P_s^i$ and $Q_t$ (Li, Swersky, and Zemel 2015; Guo, Pasunuru, and Bansal 2020). To achieve domain distribution alignment, MMD is expected to become smaller and smaller between the sampled feature distribution from the pair of source and target domains by tuning vision-aware prompts. Here, the squared MMD as defined in Eq. (11) is applied as,

$$\mathcal{L}_{DDA} = \widehat{\mathrm{MMD}}^2(\mathcal{F}, P_s^i, Q_t) \tag{12}$$

**Text Classifier Consistency.** Different domain-specific text classifiers' predictions on the edge sides are expected to be consistent when inputting the same unobserved target samples. The decentralized consistency and collaboration among these edge-side models are realized by doing so. The discrepancy among all domain-specific text classifiers is minimized by,

$$\mathcal{L}_{TCC} = \frac{1}{n_t K \mathcal{C}_M^2} \sum_{i=1}^{M-1} \sum_{r=i+1}^{M} \sum_{j=1}^{n_t} \sum_{k=1}^{K} \left| p_i(\hat{y}_{j,k}^t | x_j^t; \mathbf{P}_i) - p_r(\hat{y}_{j,k}^t | x_j^t; \mathbf{P}_r) \right| \tag{13}$$

where the combination $\mathcal{C}_M^2$ represents the number of distinct pairs that can be formed from $M$ source domains.

**Text Semantic Diversity.** Intuitively, text prompts among different domains should be diverse to describe the domain-specific semantic meaning. For example, a customized description for the sketch domain could be "`A sketch of a <category> with pencil lines.`", while the painting domain might be "`A painting of a <category> with colorful paint.`". In our implementation, the domain-agnostic text prompts are fixed as the manual-crafted prompt, i.e., `a photo of a <category>.`, and the domain-specific text prompt guided by the corresponding visual prompt is concatenated with it. It is expected that the final text prompts from different source domains are encouraged to be slightly dissimilar to better represent the domain-specific descriptions of diversity. The semantic orthogonal constraint is introduced to ensure dissimilarity as follows,

$$\begin{aligned} \mathcal{L}_{TSD} &= \frac{1}{\mathcal{C}_M^2 \cdot K} \sum_{i=1}^{M-1} \sum_{r=i+1}^{M} \sum_{k=1}^{K} \left| \cos(W_k^i, W_k^r) \right| \\ &= \frac{1}{\mathcal{C}_M^2 \cdot K} \sum_{i=1}^{M-1} \sum_{r=i+1}^{M} \sum_{k=1}^{K} \left| \cos(\Phi_i(\mathbf{E}_k, \mathbf{P}_i), \Phi_r(\mathbf{E}_k, \mathbf{P}_r)) \right| \end{aligned} \tag{14}$$

where $W^i$ represents the text representations from the $i$-th source domain.

**Training VAMP Framework.** Accordingly, the total optimization objective within each edge-side model and among edge-side models is,

$$\mathcal{L} = \mathcal{L}_{CSA} + \alpha_1 \cdot \lambda(\mathcal{L}_{DDA} + \mathcal{L}_{TCC}) + \alpha_2 \cdot \lambda \mathcal{L}_{TSD} \tag{15}$$

where $\alpha$s are the hyperparameters to balance losses. Considering that each domain-specific prompt is not fully trained at the beginning, $\lambda$ is a dynamically adjustive coefficient to control $\alpha$s, which increases with the training steps and is calculated as,

$$\lambda = 2 \cdot \mathrm{sigmoid}(10 \cdot \frac{steps}{total\_steps}) - 1 \tag{16}$$

The pseudo-training procedure of the VAMP is shown in Algorithm I, in Appendix A. During inference in the centralized integration platform, the average predicted logits from the multiple edge-side devices are used as the final prediction of target samples, which can be quickly achieved by inserting the domain-specific prompts they send.

# Experiments

This section describes the datasets, baselines, extensive experiments, and analysis. More details about datasets and implementations are shown in Appendix B and C.

## Experimental Settings

**Datasets.** VAMP is evaluated on two multi-source few-shot domain adaptation benchmarks, OfficeHome (Venkateswara et al. 2017), and DomainNet (Peng et al. 2019).

**Baselines.** We compare the VAMP with the several prompt tuning methods implemented under the UMFDA scenario and based on ViT-B/16 CLIP, which are organized into the following types.

- **Zero-shot CLIP.** It denotes that CLIP zero-shot inference on the target domain data is directly implemented.
- **Domain-agnostic Prompt.** It includes several prevalent prompt tuning methods, such as CoOp (Zhou et al. 2022b), VPT (denotes only vision-branch prompt tuning) and MaPLe (Khattak et al. 2023). In UMFDA, they are viewed as domain-specific prompts that are independently learned on each edge-side model. Only a few annotated data can be used to train. In the central model, the best inference result in the target domain is reported as the final inference result.
- **Disentangling Prompt Tuning.** It includes a single-source domain adaptation method DAPL (Ge et al. 2022) and a multi-source domain adaptation method MPA (Chen et al. 2023). Similar to the domain-agnostic prompt method mentioned above, DAPL is independently implemented on each edge-side model, and the best inference results are reported. For MPA, the reported results are from the inference after the second-stage alignment training on the central model.

## Comparative Results

Extensive experiments are conducted on OfficeHome and DomainNet datasets, as shown in Table 1. The observation and findings are as follows. (1) The zero-shot CLIP inference has achieved a good performance, which demonstrates that CLIP has the advantage of language supervision and the original text prompt is capable of inspiring the generalization knowledge of CLIP. (2) Within the domain-agnostic prompt tuning methods, MaPLe works best, demonstrating the advantage of multimodal prompt tuning to maintain the discriminability of features. (3) In the UMFDA setting, DAPL's and MPA's performances do not meet expectations. Their results reflect that decoupling prompts to capture domain-specific and domain-agnostic contexts by contrastive learning are suboptimal under inadequately annotated source domain data. (4) VAMP achieves competitive

| OfficeHome | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Method | | 3% | | | | | 6% | | | | |
| | | APR →C | CPR →A | CAR →P | CAP →R | Avg | APR →C | CPR →A | CAR →P | CAP →R | Avg |
| Zero-shot | CLIP | 67.7 | 82.7 | 89.2 | 89.6 | 82.3 | 67.7 | 82.7 | 89.2 | 89.6 | 82.3 |
| Domain-agnostic prompt | CoOP | 70.0±0.3 | 82.4±0.6 | 90.9±0.2 | 90.3±0.3 | 83.4±0.2 | 70.3±0.6 | 82.5±0.4 | 90.6±0.4 | 90.4±0.8 | 83.5±0.5 |
| | VPT | 69.6±0.4 | 83.0±0.5 | 90.2±0.2 | 90.2±0.1 | 83.3±0.2 | 69.7±0.4 | 83.9±0.6 | 90.5±0.4 | 90.3±0.4 | 83.6±0.3 |
| | MaPLe | 71.1±0.7 | 83.3±0.5 | 91.2±0.2 | 90.5±0.6 | 84.0±0.4 | 71.0±0.6 | 83.1±0.5 | **91.8**±0.4 | 90.7±0.2 | 84.1±0.3 |
| Disentangling prompt | DAPL | 70.0±0.1 | 83.5±0.7 | 91.0±0.5 | 90.5±0.2 | 83.7±0.4 | 70.2±0.3 | 84.3±0.3 | 90.9±0.4 | 90.6±0.2 | 84.0±0.1 |
| | MPA | 63.0±0.5 | 76.9±1.3 | 83.5±1.1 | 81.6±0.4 | 76.2±0.2 | 63.5±0.7 | 77.3±0.9 | 83.4±0.3 | 81.2±0.5 | 76.3±0.3 |
| VAMP | | **73.7**±0.6 | **85.7**±0.3 | **91.4**±0.4 | **90.9**±0.2 | **85.4**±0.2 | **73.5**±0.2 | **85.8**±0.4 | 91.4±0.1 | **91.4**±0.2 | **85.5**±0.1 |
| DomainNet | | | | | | | | | | | |
| Method | | 1 shot | | | | | 3 shot | | | | |
| | | PRS →C | CRS →P | CPS →R | CPR →S | Avg | PRS →C | CRS →P | CPS →R | CPR →S | Avg |
| Zero-shot | CLIP | 82.7 | 82.6 | 91.8 | 79.6 | 84.2 | 82.7 | 82.6 | 91.8 | 79.6 | 84.2 |
| Domain-agnostic prompt | CoOP | 82.3±0.6 | 81.9±0.5 | 90.9±0.4 | 79.1±0.3 | 83.5±0.4 | 83.5±0.2 | 82.8±0.3 | 91.1±0.1 | 80.0±0.6 | 84.3±0.3 |
| | VPT | 82.4±0.2 | 82.2±0.3 | 91.7±0.0 | 80.0±0.2 | 84.0±0.1 | 83.4±0.1 | 83.1±0.1 | 91.9±0.1 | 80.3±0.0 | 84.7±0.0 |
| | MaPLe | 83.3±0.6 | 82.9±0.0 | 91.8±0.3 | 79.6±0.2 | 84.4±0.1 | 84.6±0.3 | 83.3±0.2 | 91.6±0.2 | 80.1±0.5 | 84.9±0.1 |
| Disentangling prompt | DAPL | 83.4±0.8 | 83.7±0.3 | 91.9±0.4 | **80.8**±0.1 | 85.0±0.4 | 83.6±0.4 | 84.3±0.2 | 92.1±0.5 | 81.2±0.2 | 85.3±0.2 |
| | MPA | 82.5±1.1 | 82.5±0.2 | 91.4±0.1 | 80.2±0.1 | 84.2±0.2 | 83.5±0.2 | 82.6±0.2 | 91.8±0.0 | 80.4±0.3 | 84.4±0.1 |
| VAMP | | **84.3**±0.1 | **84.3**±0.1 | **92.6**±0.1 | **80.8**±0.2 | **85.5**±0.0 | **85.1**±0.1 | **84.9**±0.1 | **92.5**±0.0 | **81.9**±0.0 | **86.1**±0.0 |

Table 1: Accuracy (%) and Standard deviation (%) evaluation on target domain of OfficeHome and DomainNet dataset. A, P, R, and C in OfficeHome denote acronyms of Art, Product, Real, and Clipart, respectively. P, R, S, and C in DomainNet denote acronyms of Painting, Real, Sketch, and Clipart, respectively. Our reported results are the average of four runnings. The best results are shown in bold. Mann-Whitney U test is performed to compared our average results with the second-best average results. On OfficeHome, p-values are 0.01 for %3 and 0.01 for %6. On DomainNet, p-values are 0.02 for 1 shot and 0.01 for 3 shot. The p-values are all less than 0.05 indicating a significant difference in medians.

| OfficeHome | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Direction | 3% | | | | | 6% | | | | |
| | APR →C | CPR →A | CAR →P | CAP →R | Avg | APR →C | CPR →A | CAR →P | CAP →R | Avg |
| *vision→text* | **74.5** | **85.6** | **91.9** | 90.6 | **85.7** | **73.6** | **86.3** | **91.3** | **91.3** | **85.6** |
| *text→vision* | 73.6 | 85.4 | 91.1 | **90.7** | 85.2 | 73.3 | 85.6 | 91.0 | 91.0 | 85.2 |
| DomainNet | | | | | | | | | | |
| Direction | 1 shot | | | | | 3 shot | | | | |
| | PRS →C | CRS →P | CPS →R | CPR →S | Avg | PRS →C | CRS →P | CPS →R | CPR →S | Avg |
| *vision→text* | **84.3** | **84.3** | **92.6** | **80.9** | **85.5** | **85.0** | **84.9** | **92.5** | **81.9** | **86.1** |
| *text→vision* | 83.4 | 83.7 | 92.2 | 80.8 | 85.0 | 84.9 | 84.9 | 92.3 | 81.8 | 86.0 |

Table 2: Ablation studies on changing the direction of prompt projection in the proposed framework.

| $TSD$ | $TCC$ | $DDA$ | 3% | | | | | 6% | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | APR →C | CPR →A | CAR →P | CAP →R | Avg | APR →C | CPR →A | CAR →P | CAP →R | Avg |
| | | | **74.5** | 85.6 | **91.9** | 90.6 | **85.7** | **73.6** | **86.3** | **91.3** | 91.3 | **85.6** |
| ✗ | | | 73.7 | **85.7** | 91.7 | **90.8** | 85.5 | 73.4 | 85.6 | 91.3 | **91.5** | 85.5 |
| ✗ | ✗ | | 73.7 | 85.7 | 91.4 | 90.3 | 85.3 | 73.4 | 85.7 | 91.2 | 91.1 | 85.4 |
| ✗ | ✗ | ✗ | 74.0 | 85.2 | 91.7 | 90.3 | 85.3 | 73.5 | 85.9 | 91.2 | 91.1 | 85.4 |

Table 3: Ablation studies on various losses of VAMP on the OfficeHome dataset.

performance compared to baselines and has an average improvement of 3.2% and 1.6% compared to the zero-shot CLIP inference on the OfficeHome and DomainNet datasets, respectively. It indicates that utilizing vision-aware multimodal prompts as domain-specific prompts to perceive domain information alleviates the limitations of disentangling prompts in the few-shot setting. Further, it demonstrates that

| b | 3% | | | | | 6% | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | APR →C | CPR →A | CAR →P | CAP →R | Avg | APR →C | CPR →A | CAR →P | CAP →R | Avg |
| 2 | 73.2 | 85.6 | 91.5 | 91.0 | 85.3 | 72.7 | 85.5 | 91.5 | 91.6 | 85.3 |
| 4 | 73.2 | 85.2 | 91.4 | 91.0 | 85.2 | 73.4 | 85.8 | 91.5 | 91.4 | 85.5 |
| 8 | 73.4 | 85.5 | 91.6 | 91.2 | 85.4 | 72.9 | 86.1 | 91.4 | 91.4 | 85.5 |
| 16 | 73.7 | 85.9 | 91.0 | 90.9 | 85.4 | 73.5 | 85.5 | 91.5 | 91.4 | 85.5 |

Table 4: Ablation studies on prompt lengths $b$ of VAMP on OfficeHome dataset.

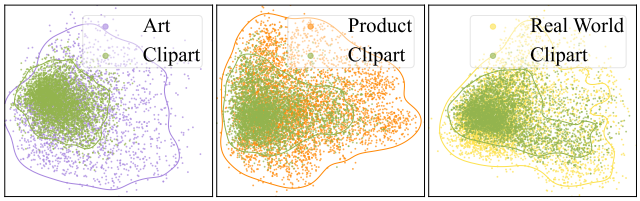
this decentralized training framework of the VAMP can simultaneously address the issues of domain and semantic alignment within edge-side models and collaborative learning among edge-side models without central training.

## Ablation Studies

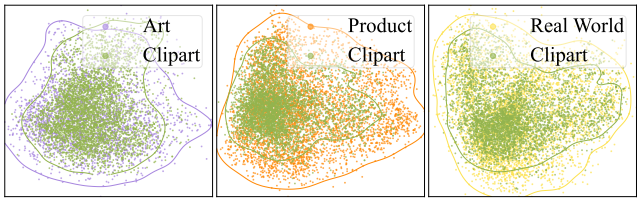
**Effects of Vision-aware Multimodal Prompts.** Table 2 shows the effectiveness of vision-aware multimodal prompts in projecting direction from vision to text (*vision→text*). It demonstrates that this multimodal prompt tuning is better for perceiving domain information from the vision encoder and affects the text prompt tuning.

**Effects of Various Losses.** Table 3 shows the declines of various degrees for the VAMP framework when various losses are removed on OfficeHome. It validates the effectiveness of these introduced losses to jointly optimize the VAMP better.

**Effects of Prompt Length.** Table 4 shows the effects of prompt length for the vision-aware multimodal prompts on OfficeHome. The ablation results indicate that increasing its




(a) *text→vision*




(b) *vision→text*

Figure 4: PCA visualizations of the extracted image features from the source and target domains in different domain-specific models of "APR→C". The first row of pictures denotes that the projection direction of prompts is from *text* to *vision*; the second row represents the *vision*-to-*text* projection that is used in the VAMP. Contour lines enclose the regions with high density of data points.

prompt length has a limited impact on performance.

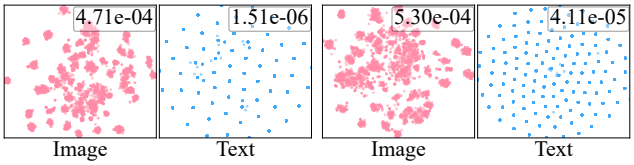
## Quantity Analysis

**Domain Distribution Visualization.** Figure 4 depicts the domain distribution visualization in two projection directions (*text→vision* and *vision→text*). The *vision→text* projection has a better alignment between the source domain and target domain in each domain-specific image extractor, demonstrating the effectiveness of vision-aware multimodal prompt to perceive the domain information.

**t-SNE Visualization and Variance Statistics.** Figure 5 shows the different t-SNE visualization and variance statistics of the image and text features from target domain. The observation is that the image features of VAMP are more concentrated aggregations within classes, and it has fewer intra-class visual variances. The text feature distribution of VAMP between classes is not even as DAPL. Because of only text-branch prompt tuning, DAPL has a more significant inter-class text variance, showing more explicit inter-class boundaries. Although there is a loss in the inter-class distribution of text features, it demonstrates that our VAMP can balance feature discriminability and domain alignment by vision-aware prompt influencing text prompt. Case studies about visualizing attention maps are presented in Appendix D.

## Related Work

**MFDA.** MFDA is a more common setting (Gulshan et al. 2016; Harmon et al. 2020) since it only has fewer annotated source samples. The first insight into MFDA comes from MSFAN (Yue et al. 2021a). Like most conventional multi-source domain adaptation (MDA) methods (Venkat et al. 2020; Peng et al. 2019; Xu et al. 2018), its framework comprises a shared feature and multiple domain-specific classifiers. MSFAN adopts similarity-based classifiers (Saito et al. 2019), which are trained by prototypical contrastive learning (Yue et al. 2021b) in each pair of source and target domains to learn the well-clustered prototype features with better semantic discriminability, instead of using the linear classifiers. This paper further considers a UMFDA setting suited for the resource-limited edge device. More comparisons of our VAMP in the MFDA are presented in Appendix E.




(a) VAMP (b) DAPL

Figure 5: t-SNE visualization of the image and text features of target domain extracted by "Clipart-Real World" model of VAMP and DAPL. The statistics of either intra-class visual variance or inter-class text variance are shown at the top of the subfigure.

**Prompt Tuning in Domain Adaptation.** Prompt tuning technology has efficiently adapted pretrained models to various domains (Jia et al. 2022; Yao et al. 2021; Jin et al. 2022; Ju et al. 2022). Some domain adaptation prompt tuning methods have emerged recently (Ge et al. 2022; Chen et al. 2023; Singha et al. 2023; Bai et al. 2024; Wang et al. 2024). DAPL (Ge et al. 2022) pioneers the application of prompt tuning in single-source unsupervised domain adaptation. It disentangles context prompts as domain-agnostic and domain-specific prompts, where the domain information is shared by the same domain and thus dynamically adapts to the text classifier encoder. Subsequently, MPA (Chen et al. 2023) continually develops the idea of disentangling prompts in the MDA. After learning individual prompts for each pair of source and target domains, they excavate the relations among the learned prompts by the auto-encoder networks. Differently, this paper explores the prompt tuning in MDA under a resource-limited scenario with limited data.

## Conclusions

This study introduces a UMFDA schema for the resource-limited edge computing scenario, inspired by the advantage of pretrained VLM and the plug-and-play prompt tuning capable of efficiently transferring the VLM. We further propose a decentralized training VAMP framework. It is based on the customized vision-aware multimodal prompts to perceive the domain information and maintain the features' discriminability. Within the VAMP framework, CSA, DDA, TCC, and TSD losses are introduced to optimize the alignment inside the edge-side model and to enhance collaboration among edge-side models. The experimental results demonstrate the VAMP's effectiveness. Future work will explore VAMP's performance on more edge-side models in the various downstream tasks of MFDA.

## Acknowledgements

This work was supported by the National Natural Science Foundation of China (NSFC) under Grant Nos.61966038, 62266051, 62202416 and 62162068, and the Postgraduate Research and Innovation Foundation of Yunnan University under Grant No.KC-24248816. The authors would like to thank the anonymous reviewers for their constructive comments.

## References

Bai, S.; Zhang, M.; Zhou, W.; Huang, S.; Luan, Z.; Wang, D.; and Chen, B. 2024. Prompt-Based Distribution Alignment for Unsupervised Domain Adaptation. In *Proceedings of the 38th AAAI Conference on Artificial Intelligence (AAAI-2024), the 36th Conference on Innovative Applications of Artificial Intelligence, (IAAI-2024), the 14th Symposium on Educational Advances in Artificial Intelligence, (EAAI-2024)*, volume 38, 729–737.

Chen, G.; Yao, W.; Song, X.; Li, X.; Rao, Y.; and Zhang, K. 2022. PLOT: Prompt Learning with Optimal Transport for Vision-Language Models. arXiv:2210.01253.

Chen, H.; Han, X.; Wu, Z.; and Jiang, Y.-G. 2023. Multi-Prompt Alignment for Multi-Source Unsupervised Domain Adaptation. In *Advances in Neural Information Processing Systems (NeurIPS-2023)*, 74127–74139.

Du, Z.; Li, X.; Li, F.; Lu, K.; Zhu, L.; and Li, J. 2024. Domain-Agnostic Mutual Prompting for Unsupervised Domain Adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR-2024)*, 23375–23384.

Ge, C.; Huang, R.; Xie, M.; Lai, Z.; Song, S.; Li, S.; and Huang, G. 2022. Domain Adaptation via Prompt Learning. arXiv:2202.06687.

Gretton, A.; Borgwardt, K. M.; Rasch, M.; Schölkopf, B.; and Smola, A. J. 2007. A Kernel Method for the Two-Sample-Problem. In *Advances in Neural Information Processing Systems (NeurIPS-2006)*, 513–520.

Gretton, A.; Borgwardt, K. M.; Rasch, M.; Schölkopf, B.; and Smola, A. J. 2012. A Kernel Two-Sample Test. *Journal of Machine Learning Research*, 13: 723–773.

Gulshan, V.; Peng, L.; Coram, M.; Stumpe, M. C.; Wu, D.; Narayanaswamy, A.; Venugopalan, S.; Widner, K.; Madams, T.; Cuadros, J.; Kim, R.; Raman, R.; Nelson, P. C.; Mega, J. L.; and Webster, D. R. 2016. Development and Validation of a Deep Learning Algorithm for Detection of Diabetic Retinopathy in Retinal Fundus Photographs. *JAMA*, 316(22): 2402.

Guo, H.; Pasunuru, R.; and Bansal, M. 2020. Multi-Source Domain Adaptation for Text Classification via DistanceNet-Bandits. *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI-2020), the 32nd Innovative Applications of Artificial Intelligence Conference (IAAI-2020), the 10th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-2020)*, 34(05): 7830–7838.

Harmon, S. A.; Sanford, T. H.; Xu, S.; Turkbey, E. B.; Roth, H.; Xu, Z.; Yang, D.; Myronenko, A.; Anderson, V.; Amalou, A.; Blain, M.; Kassin, M.; Long, D.; Varble, N.; Walker, S. M.; Bagci, U.; Ierardi, A. M.; Stellato, E.; Plensich, G. G.; Franceschelli, G.; Girlando, C.; Irmici, G.; Labella, D.; Hammoud, D.; Malayeri, A.; Jones, E.; Summers, R. M.; Choyke, P. L.; Xu, D.; Flores, M.; Tamura, K.; Obinata, H.; Mori, H.; Patella, F.; Cariati, M.; Carrafiello, G.; An, P.; Wood, B. J.; and Turkbey, B. 2020. Artificial intelligence for the detection of COVID-19 pneumonia on chest CT using multinational datasets. *Nature Communications*, 11(1): 4080.

Jia, M.; Tang, L.; Chen, B. C.; Cardie, C.; Belongie, S.; Hariharan, B.; and Lim, S. N. 2022. Visual Prompt Tuning. In *Proceedings of the 17th European Conference on Computer Vision (ECCV-2022)*, volume 13693, 709–727.

Jin, W.; Cheng, Y.; Shen, Y.; Chen, W.; and Ren, X. 2022. A Good Prompt Is Worth Millions of Parameters: Low-resource Prompt-based Learning for Vision-Language Models. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (ACL-2022)*, 2763–2775.

Ju, C.; Han, T.; Zheng, K.; Zhang, Y.; and Xie, W. 2022. Prompting Visual-Language Models for Efficient Video Understanding. In *Proceedings of the 17th European on Conference Computer Vision (2022-ECCV)*, 105–124.

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 (CVPR-2023)*, 19113–19122.

Kim, D.; Saito, K.; Oh, T.-H.; Plummer, B. A.; Sclaroff, S.; and Saenko, K. 2020. Cross-domain Self-supervised Learning for Domain Adaptation with Few Source Labels. arXiv:2003.08264.

Lester, B.; Al-Rfou, R.; and Constant, N. 2021. The Power of Scale for Parameter-Efficient Prompt Tuning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP-2021)*, 3045–3059.

Li, X. L.; and Liang, P. 2021. Prefix-Tuning: Optimizing Continuous Prompts for Generation. arXiv:2101.00190.

Li, Y.; Swersky, K.; and Zemel, R. 2015. Generative moment matching networks. arXiv:1502.02761.

Liu, X.; Ji, K.; Fu, Y.; Tam, W. L.; Du, Z.; Yang, Z.; and Tang, J. 2021. P-Tuning v2: Prompt Tuning Can Be Comparable to Fine-tuning Universally Across Scales and Tasks. arXiv:2110.07602.

Lu, Y.; Liu, J.; Zhang, Y.; Liu, Y.; and Tian, X. 2022. Prompt Distribution Learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR-2022)*, 5196–5205.

McMahan, H. B.; Moore, E.; Ramage, D.; Hampson, S.; and Arcas, B. A. y. 2016. Communication-Efficient Learning of Deep Networks from Decentralized Data. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics (AISTATS-2017)*, 1273–1282.

Peng, X.; Bai, Q.; Xia, X.; Huang, Z.; Saenko, K.; and Wang, B. 2019. Moment Matching for Multi-Source Domain Adaptation. In *Proceedings of the IEEE/CVF Interna-*

*tional Conference on Computer Vision (ICCV-2019)*, 1406–1415.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; Krueger, G.; and Sutskever, I. 2021. Learning Transferable Visual Models from Natural Language Supervision. In *Proceedings of the 38th International Conference on Machine Learning (ICML-2021)*, 8748–8763.

Saito, K.; Kim, D.; Sclaroff, S.; Darrell, T.; and Saenko, K. 2019. Semi-Supervised Domain Adaptation via Minimax Entropy. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV-2019)*, 8049–8057.

Shin, T.; Razeghi, Y.; Logan IV, R. L.; Wallace, E.; and Singh, S. 2020. AutoPrompt: Eliciting Knowledge from Language Models with Automatically Generated Prompts. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP-2020)*, 4222–4235.

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 Workshops (ICCV Workshops-2023)*, 4357–4366.

Venkat, N.; Kundu, J. N.; Singh, D. K.; Revanur, A.; and Babu, R. V. 2020. Your Classifier can Secretly Suffice Multi-Source Domain Adaptation. In *Advances in Neural Information Processing Systems (NeurIPS-2020)*, 4647–4659.

Venkateswara, H.; Eusebio, J.; Chakraborty, S.; and Panchanathan, S. 2017. Deep Hashing Network for Unsupervised Domain Adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR-2017)*, 5385–5394.

Wang, Z.; Zhang, L.; Wang, L.; and Zhu, M. 2024. LanDA: Language-Guided Multi-Source Domain Adaptation. arXiv:2401.14148.

Xu, R.; Chen, Z.; Zuo, W.; Yan, J.; and Lin, L. 2018. Deep Cocktail Network: Multi-source Unsupervised Domain Adaptation with Category Shift. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR-2018)*, 3964–3973.

Yao, Y.; Zhang, A.; Zhang, Z.; Liu, Z.; Chua, T.-S.; and Sun, M. 2021. CPT: Colorful Prompt Tuning for Pre-trained Vision-Language Models. arXiv:2109.11797.

Yue, X.; Zheng, Z.; Reed, C.; Das, H. P.; Keutzer, K.; and Vincentelli, A. S. 2021a. Multi-source Few-shot Domain Adaptation. arXiv:2109.12391.

Yue, X.; Zheng, Z.; Zhang, S.; Gao, Y.; Darrell, T.; Keutzer, K.; and Vincentelli, A. S. 2021b. Prototypical Cross-domain Self-supervised Learning for Few-shot Unsupervised Domain Adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR-2021)*, 13829–13839.

Zang, Y.; Li, W.; Zhou, K.; Huang, C.; and Loy, C. C. 2022. Unified Vision and Language Prompt Learning. arXiv:2210.07225.

Zhao, Z.; Gan, L.; Wang, G.; Hu, Y.; Shen, T.; Yang, H.; Kuang, K.; and Wu, F. 2024. Retrieval-Augmented Mixture of LoRA Experts for Uploadable Machine Learning. arXiv:2406.16989.

Zhou, K.; Yang, J.; Loy, C. C.; and Liu, Z. 2022a. Conditional Prompt Learning for Vision-Language Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR-2022)*, 16795–16804.

Zhou, K.; Yang, J.; Loy, C. C.; and Liu, Z. 2022b. Learning to Prompt for Vision-Language Models. *International Journal of Computer Vision*, 130(9): 2337–2348.

Zhu, Y.; Zhuang, F.; and Wang, D. 2019. Aligning Domain-Specific Distribution and Classifier for Cross-Domain Classification from Multiple Sources. In *Proceedings of the 33rd AAAI Conference on Artificial Intelligence (2019-AAAI)*, volume 33, 5989–5996.