AD-CLIP: Adapting Domains in Prompt Space Using CLIP

Mainak Singha* Harsh Pal* Ankit Jha* Biplab Banerjee
Indian Institute of Technology Bombay, India
{mainaksingha.iitb, palharsh.india, ankitjha16, getbiplab}@gmail.com

Abstract

Although deep learning models have shown impressive performance on supervised learning tasks, they often struggle to generalize well when the training (source) and test (target) domains differ. Unsupervised domain adaptation (DA) has emerged as a popular solution to this problem. However, current DA techniques rely on visual backbones, which may lack semantic richness. Despite the potential of large-scale vision-language foundation models like CLIP, their effectiveness for DA has yet to be fully explored. To address this gap, we introduce AD-CLIP*, a domain-agnostic prompt learning strategy for CLIP that aims to solve the DA problem in the prompt space. We leverage the frozen vision backbone of CLIP to extract both image style (domain) and content information, which we apply to learn prompt tokens. Our prompts are designed to be domain-invariant and class-generalizable, by conditioning prompt learning on image style and content features simultaneously. We use standard supervised contrastive learning in the source domain, while proposing an entropy minimization strategy to align domains in the embedding space given the target domain data. We also consider a scenario where only target domain samples are available during testing, without any source domain data, and propose a cross-domain style mapping network to hallucinate domain-agnostic tokens. Our extensive experiments on three benchmark DA datasets demonstrate the effectiveness of* AD-CLIP *compared to existing literature.*

1. Introduction

The use of deep convolutional neural networks (convnets) has led to significant advancements in visual recognition tasks within supervised learning settings [16]. These models can learn discriminative, data-driven features from large sets of training samples. However, they are vulnerable to the domain-shift problem, which arises when training and test samples come from different distributions, causing the probable approximately correct (PAC) [9] assumption to fail. One potential solution to this issue is domain adaptation (DA) [41, 4, 1], a form of transductive transfer learning. DA leverages both labeled source data and unlabeled target data to create a domain-agnostic embedding space. This space can be used to train classifiers on the source domain, which can accurately classify target samples. Numerous DA techniques are available in the literature, including adversarial, entropy minimization, and statistical distance optimization approaches [37]. *However, current models typically rely on convnets [10] purely trained on visual data, which often lack semantic richness, thus producing sub-optimal performance in critical situations* (see Fig. 1).

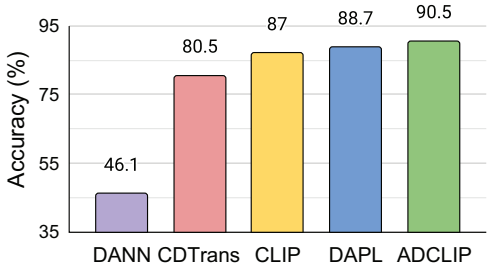

Figure 1: We compare the performance of AD-CLIP for the Office-Home [36] dataset, with different type of UDA methods, e.g. convnets-based DANN [6], Transformer-based CDTrans [39], pre-trained CLIP [26] without prompt learning and, DAPL [7], a prompt learning-based DA technique.

Large-scale vision-language models [26, 13] have emerged as the *de-facto* feature extractor in computer vision nowadays. These models are trained on vast quantities of image-text pairs, where class labels are represented as textual prompts (e.g., `a photo of a [CLS]`). This results in a joint embedding space with rich semantics, facilitating excellent generalization and zero-shot classification performance. Although the CLIP [26] model has demonstrated impressive results, designing task-specific prompts can be challenging. Subsequent works [44, 43] have focused on learning prompts in a data-driven manner, primarily utilizing visual information from CLIP's frozen image

*equal contribution

backbone. Despite the success of these approaches, they have yet to be meaningfully applied for cross-domain inference tasks, with only a few works focusing on domain generalization [2]. *In opposition, our focus is on solving the DA problem by leveraging the semantic richness of CLIP. To achieve this goal, we aim to use the pre-trained backbones of CLIP without fine-tuning, but propose to learn prompts that can capture the domain and class distributions well and are domain-agnostic, by introducing only a small set of learnable parameters.*

Our proposed AD-CLIP: In this paper, we introduce a novel framework called AD-CLIP to address our research questions. Our main objective is to design prompts that can generalize well across the source and the target domains. To achieve this, we propose a method to learn new prompts that consist of two types of tokens: i) *Domain token*: To incorporate domain knowledge, we introduce a token that captures the style information of both domains. Style corresponds to the feature statistics obtained from CLIP's image encoder [20], and we propose a way to combine the multi-scale style features through a *style projector*. ii) *Image-specific tokens*: To learn the visual distributions well in the semantic space and obtain a distribution of prompts per class, we leverage the visual feature responses from the different layers of CLIP's vision encoder to initiate learning of image-conditioned tokens. We note that we consider the multi-scale features in this aspect, as they can better characterize the underlying visual concepts at multiple abstractions. We introduce a set of *content projectors* for this purpose, and all the projectors are trained contrastively given the prompt and image embeddings. For aligning the target domain data with the source counterparts, we propose simple yet effective entropy minimization characteristics given the similarity distributions between the image embeddings and the prompt embeddings for all the classes while aligning the cross-domain prompt embeddings through optimizing a measure of distributions divergence.

During the inference stage, we often only have access to the test images from the target domain, without any access to the source domain data. This creates a challenge when trying to define the domain-driven prompt token without the knowledge of the source domain, which is only available during training. To overcome this, we propose a novel approach that involves hallucinating the source domain characteristics based on the target domain properties, using a *target-to-source style prediction network*. We train this network by passing the style features of the target domain images, which are then used to generate the corresponding source style information. Our approach distinguishes itself from other prompt learning methods described in the literature [44, 43] in the way we propose to generate the prompts while ensuring domain independence and generalizability jointly (see Figure 2). Our **significant contributions** are summarized as follows:

[-] We propose a solution to the challenging domain adaptation problem using prompt learning within the context of CLIP. Our primary focus is to ensure that the prompts are not biased towards a specific domain and account for the visual variations in the data.

[-] To achieve this, we propose a novel prompt learning scheme that entirely leverages the visual encoder of CLIP and introduces a small set of learnable projector networks. We also propose a new entropy minimization-based criterion for domain alignment. Furthermore, we address the scenario where source domain data are not available during inference and develop a method to approximate the prompts for the target images.

[-] Through extensive experiments on three widely-used benchmark DA datasets, namely Office-Home [36], VisDA [25], and mini-DomainNet [24], we demonstrate the superior performance of AD-CLIP over state-of-the-art alternatives.

2. Related Works

Unsupervised Domain Adaptation: DA is the process of adapting a machine learning model trained on a source domain to a target domain where the data distributions may differ. The literature is rich with a plethora of DA approaches, including distribution alignment based on sub-space alignment, pseudo-labeling, or adversarial techniques, among others. For example, Maximum Mean Discrepancy (MMD) [17] reduces the distance between the distributions of the source and target domains in the kernel space. Another popular approach is DANN [6], which involves adding a domain classifier to the deep neural network, enabling it to learn to distinguish between source and target domain data. CyCADA [11] utilizes cycle-consistent adversarial learning to align the feature distributions. CDTrans [39] uses cross-attention and two-way center-aware labeling in Transformers [35] for domain alignment, making it robust to noisy label pairs. A more detailed discussion on DA can be found in [37, 42]. Recent approaches have started considering vision-language models for solving the DA task given their enhanced feature space. *However, the existing sole method in this regard, DAPL [7], uses ad-hoc prompting to learn disentangled domain and category representations. DAPL [7] manually includes the domain information, which is unrealistic in some cases. Additionally, DAPL [7] ignores the visual distributions of the classes, causing overfitting.*

Vision-Language models and Prompt Learning: The large-scale vision-language models (VLMs) integrate visual and textual inputs to achieve a more comprehensive understanding of the world, leading to better performance in various computer vision tasks. They typically rely on

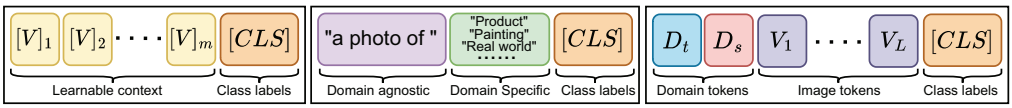

Figure 2: We highlight the differences between our prompts from the literature. a) CoOp [44] directly learns the prompt tokens from random vectors and may not be suitable for DA as it does not concern domain independence, b) Another possibility is to manually include the `domain name` into manually defined prompts, but this information may not be readily available, c) AD-CLIP introduces an automatic solution by leveraging the visual space to define the domain-agnostic and image-conditioned tokens.

pre-trained language models, such as BERT [5] and GPT [27], to encode textual inputs, while the visual inputs are processed using convnets or vision transformers. Some of the popular VLMs are CLIP [26] and VisualBERT [18].

In a similar spirit, prompt learning for VLMs is a technique that has gained increasing attention in computer vision, which involves leveraging pre-trained language models to provide valuable insights for downstream tasks through prompts. Several recent studies have explored the use of prompt learning, such as CoOp [44], and CoCoOp [43], which use conditional prompts to improve the model's generalization capabilities. AutoPrompt [29] explores tokens with the most significant gradient changes in the label likelihood to automate the prompt generation process. Whereas, APPLeNet [30] addresses the problem of DG in remote sensing by introducing prompt learning. Another recent study, MaPLe [14], proposes multi-modal prompt learning to avoid possible overfitting. *However, none of the existing prompting techniques is tailored for the DA task except DAPL [7], which is majorly hand-crafted. In contrast, we propose a more robust prompt learning technique while ensuring domain independence and improving the adaptation capabilities both in image and text feature space.*

3. Proposed Methodology

Problem Definition: The DA problem involves a source domain with the image-label pairs, $\mathcal{D}^{\mathcal{S}_l} = \{x_i^{\mathcal{S}_l}, y_i^{\mathcal{S}_l}\}_{i=1}^{N_{\mathcal{S}_l}}$ ($x_i \in \mathcal{X}^s$, $y_i \in \mathcal{Y}$), where the labeled data follows the joint distribution $\mathcal{P}_{data}^{\mathcal{S}_l}$, and a target domain with unlabeled images, $\mathcal{D}^{\mathcal{T}_u} = \{x_j^{\mathcal{T}_u}\}_{j=1}^{N_{\mathcal{T}_u}}$, where the unlabeled data follows the distribution $\mathcal{P}_{data}^{\mathcal{T}_u}$, respectively. It is important to note that $\mathcal{P}_{data}^{\mathcal{T}_u}$ is not equal to $\mathcal{P}_{data}^{\mathcal{S}_l}$, leading to domain shift. The number of images in the source and target domains is denoted by $N_{\mathcal{S}_l}$ and $N_{\mathcal{T}_u}$, respectively. Also, in the closed-set approach that we follow, $\mathcal{S}_l$ and $\mathcal{T}_u$ share the same label space $\mathcal{Y}$. Under this setting, the objective is to learn a classifier $f : \mathcal{X}^s \rightarrow \mathcal{Y}$ that performs well on $\mathcal{T}_u$ by leveraging $\mathcal{S}_l$ and $\mathcal{T}_u$, which requires overcoming the distributional differences between $\mathcal{D}^{\mathcal{S}_l}$ and $\mathcal{D}^{\mathcal{T}_u}$.

Overview of AD-CLIP: In the following, we delve into the details of AD-CLIP. Our primary goal is to learn domain- and class-agnostic prompts that lead to a discriminative and domain-aligned semantic embedding space. To achieve this, we utilize the frozen vision and text backbones of CLIP, referred to as f_v and f_t, respectively, both of which rely on transformers. To enable the learning of prompt tokens using visual information from different layers of f_v, we introduce learnable style and content projectors, P_v and C_v, respectively. Specifically, given f_v comprising M encoder layers, P_v and C_v facilitate prompt learning in parallel by separately looking into the image domain and content properties. Furthermore, we incorporate the target-to-source style mapping unit P_{smn} to hallucinate source style features from the target domain samples during inference. While P_{smn} and P_v take the form of an encoder-decoder, C_v is designed to consist of a single encoder and L decoders, one per prompt token, where L is the context length for the prompts.

We proceed to discuss the following: i) prompt learning in AD-CLIP using disentangled visual style and content information, ii) the target-to-source style mapping network, and iii) the loss functions for classification and domain alignment.

- Our proposed prompt learning: Our objective is to learn prompts directly from the visual domain to effectively encode the visual distribution, as opposed to the static prompting technique [44]. In this regard, we have two primary objectives for addressing the DA task: i) incorporating a domain-agnostic token into the prompt to prevent domain bias, and ii) enhancing the learning of visual concepts in prompt tokens by utilizing feature responses from multiple layers of the CLIP vision encoder. We introduce a domain-agnostic token of the form $[D_s; D_t]$. To obtain D_s, we pass the multi-scale style information through the shared style projector P_v. Precisely, the style information is represented by the first and second-order batch-wise feature statistics: $[\mu, \sigma]$. In our case, we calculate and combine $[\mu_1; \sigma_1; \cdots; \mu_M; \sigma_M]$ from the M layers of f_v for a given x to obtain the style vector $\bar{\mathcal{F}}(x)$. Similarly, we

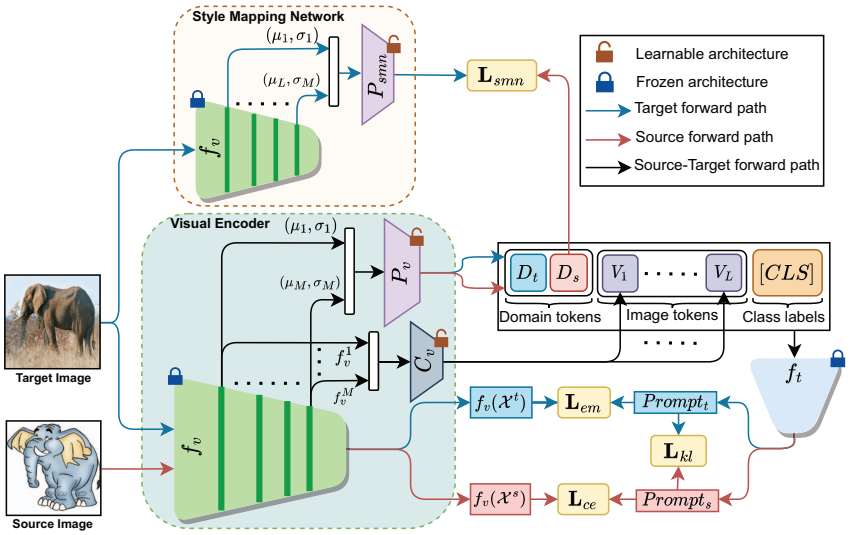

Figure 3: **The architecture of AD-CLIP** is based on the frozen CLIP backbones f_v and f_t. For prompt token learning, we introduce the new vision and text projectors P_v and C_v, respectively, which encoder the style and content information from the different layers of f_v. The style mapping network, P_{smn}, approximates the source domain style information from the target domain features. Best viewed in color.

define the multi-scale image content features as $\hat{\mathcal{F}}(x) = [\hat{f}_v^1(x); \cdots; \hat{f}_v^M(x)]$ where f_v^m denotes the responses from the m^{th} layer. For a given batch, we consider the samples from $\mathcal{S}_l$ to produce $\bar{\mathcal{F}}_s$ and samples from $\mathcal{T}_u$ to produce $\bar{\mathcal{F}}_t$, respectively, for each x. P_v subsequently maps $\bar{\mathcal{F}}_s$ onto D_s and $\bar{\mathcal{F}}_t$ onto D_t. On the other hand, $\hat{\mathcal{F}}(x)$ is passed through C_v to produce L image-specific context tokens $\{V_l\}_{l=1}^{L}$. Finally, we denote the prompt for a class y with the class embedding CLS_y given an image x as,

$$\text{Prompt}_y(x) = \Big[[D_t; D_s]; V_1; \cdots; V_L; [CLS_y]\Big] \quad (1)$$

- The target to source style mapping network: Although the model is trained in a transductive manner, where samples from both domains are used during training, it may encounter target domain samples separately during the inference stage. This absence of source domain samples during inference can hinder the generation of prompts, as the prompt relies on the domain-agnostic token. To address this challenge, we introduce a cross-domain style mapping network, denoted as f_{smn}. This network takes $\bar{\mathcal{F}}_t^b$ as input and learns to produce the corresponding D_s given the samples for the b^{th} batch. We train f_{smn} using the ℓ_2 loss, defined as follows:

$$\mathbf{L}_{smn} = \underset{f_{smn}, P_v}{\arg\min} \; \mathbb{E}_{P_{data}^{\mathcal{S}_l}, P_{data}^{\mathcal{T}_u}} ||D_s - f_{smn}(\bar{\mathcal{F}}_t)||_2^2 \quad (2)$$

- Loss functions pertaining to domain alignment: We consider two loss objectives for learning the classifier for $\mathcal{S}_l$, while aligning the prompts for $\mathcal{T}_u$ with those of the source domain. The source domain supervised contrastive loss between $f_v(x)$ and Prompt_y given (x, y) is used for image-text mapping and is optimized through the cross-entropy loss $\mathbf{L}_{ce}$. In this regard, the prediction probability of x for label y is defined as,

$$p(y|x) = \frac{\exp(\text{sim}(f_v(x), f_t(\text{Prompt}_y(x)))/\tau)}{\sum_{k=1}^{|\mathcal{Y}|} \exp(\text{sim}(f_v(x), f_t(\text{Prompt}_{y_k}(x)))/\tau)} \quad (3)$$

where, 'sim' denotes the `cosine` similarity, and τ is the temperature hyper-parameter.

Ideally, we can treat the prompts as class prototypes and aim to map the visual features onto these prototypes. This means we aim to ensure that target domain samples are aligned to one prototype while being pushed away from others to achieve domain alignment. Additionally, we want to increase the correlation between the prompts generated from the two domains, since we seek to generate a prompt-aligned semantic space. To accomplish both tasks, we have introduced a new loss objective $\mathbf{L}_{Align}$. Our approach minimises the distribution divergence between the source and target prompt embeddings using a Kullback-Leibler (KL) divergence loss ($\mathbf{L}_{KL}$). We also constrain the similarity distribution between the visual features of the target samples and prompt embeddings to have low entropy through $\mathbf{L}_{em}$. Together, they enforce the model to produce similar types of prompt embedding distributions for both domains, while aligning each target sample to a single prompt in a discriminative fashion. 'Prompt_s' denotes all the prompts

in the source domain and likewise for t.

$$\mathbf{L}_{Align} = \arg\min_{P_v, C_v} \mathop{\mathbb{E}}_{(x,y)\in P_{data}^{\mathcal{S}_l}} \mathbf{L}_{em}([p(y_1|x); \cdots; p(y_{|\mathcal{Y}|}|x)]) + \mathbf{L}_{KL}(\text{Prompt}_t|\text{Prompt}_s) \tag{4}$$

- Total loss: We train AD-CLIP with respect all the losses mentioned above as: $\mathbf{L}_{total} = [\mathbf{L}_{ce} + \mathbf{L}_{smn} + \mathbf{L}_{Align}]$.

Inference involves comparing the embeddings of the target samples to all the class prompt embeddings and selecting the class maximizing $p(y|x)$.

4. Experimental Evaluations

Datasets descriptions: We validate our model on three publicly available DA datasets. i) **Office-Home** [36]: This dataset is comprised of 15,500 high-quality images from four distinct domains: Art (Ar), Clip Art (Cl), Product (Pr), and Real World (Rw). Each domain contains a diverse range of objects from 65 different categories, set within both office and home environments. ii) **VisDA-2017** [25]: The VisDA-2017 dataset presents a more challenging scenario for synthetic-to-real domain adaptation, featuring 12 categories with 152,397 synthetic images generated by rendering 3D models from different angles and light conditions, and 55,388 real-world images collected from MSCOCO. To maintain consistency with established protocols [21, 28], we use the synthetic images as the source domain and the real-world images as the target domain. iii) **Mini-domainNet**: Lastly, we consider a subset of the comprehensive DomainNet dataset [24] called Mini-DomainNet. This subset features four domains, including Clipart (c), Painting (p), Real (r), and Sketch (s), each with images from 126 categories.

Architecture Details: For our experiments, we utilize three pre-trained vision encoders as f_v: ResNet-50 (RN50) [10], ViT-L/14, and ViT-B/16 [8] for validation. Meanwhile, we employ a transformer-based text encoder as f_t. To facilitate our projective transformation, we implement the P_v and P_{smn} projector networks using a single encoder and decoder layer. On the other hand, C_v consists of a dense encoder and L dense decoder layers, respectively.

Training and evaluation protocols: We optimize $\mathbf{L}_{total}$ using the *Adam* [15] optimizer, given a mini-batch size of 16, and an initial learning rate of 0.01, respectively. Finally, we report the target domain top-1 accuracy (`mean ± std.`) over three runs as the evaluation metric. We compare AD-CLIP against traditional DA-techniques based on vision backbones like ResNet-50 [6, 12, 38, 32], pre-trained CLIP and DAPL [7] visual features based on ViT-B/16 [8] and ViT-L/14 [8], to name a few.

4.1. Comparisons to the state-of-the-art

In this section, we present the results of our extensive evaluation of AD-CLIP alongside several state-of-the-art methods for domain adaptation (DA) on three benchmark datasets: Office-Home, VisDA-2017, and Mini-DomainNet. We also compare AD-CLIP with traditional CNN-based and Transformer-based unsupervised domain adaptation (UDA) methods, as well as vision-language foundation models. The evaluation results, presented in Tables 1-3, demonstrate that AD-CLIP achieved substantial improvements on all three benchmark datasets. Specifically, it surpassed the prior best by 1.8% in Office-Home [36], by 2.2% in VisDA-2017 [25], and by 1.6% in Mini-DomainNet [24], thereby establishing a new performance benchmark for domain adaptation tasks. In comparison with traditional CNN-based and Transformer-based UDA methods, along with vision-language foundation models, AD-CLIP consistently exhibited superior performance. Notably, AD-CLIP outperformed these methods across various evaluation metrics, reaffirming its effectiveness as a robust solution for domain adaptation challenges.

However, we have observed that on the VisDA-2017 dataset, AD-CLIP was not able to outperform the best results from different models on 7 out of 12 classes when ResNet-101 was used as the vision encoder backbone. In particular, traditional Transformer-based methods (SSRT and CDTrans) achieved the overall best results on 4 classes, namely *car*, *knife*, *person*, and *plant*. Nevertheless, the average performance of AD-CLIP still outperformed DAPL by 0.8%, 1.2%, and 2.2% when using ResNet-101, ViT-B/16, and ViT-L/14 as backbones, respectively. On the Mini-DomainNet dataset, we utilized CLIP and DAPL as baselines for comparison with AD-CLIP. The results presented in Table 3 illustrate that AD-CLIP outperformed both baseline methods by a considerable margin across all backbone models. The comprehensive evaluation indicates that AD-CLIP is a competitive method for domain adaptation tasks, achieving notable performance improvements on multiple benchmark datasets. While its performance on some classes of the VisDA-2017 dataset exhibited slight limitations, AD-CLIP remains a powerful approach, consistently outperforming existing methods on diverse datasets and backbone configurations. The results on Mini-DomainNet further validate AD-CLIP's efficacy as a reliable choice for addressing domain shift challenges. Overall, the outcomes underscore the potential of AD-CLIP as an effective and versatile tool in the domain adaptation landscape.

4.2. Ablation analysis

t-SNE [34] visualization: In this section, we conduct an ablation study to gain insights into the domain invariance and discriminativeness achieved by our proposed model, AD-CLIP. Specifically, we visualize the t-SNE representations

Table 1: Comparison of AD-CLIP with state-of-the-art methods for UDA task on Office-Home [36] dataset. We show our results with three different vision backbones; ResNet50[10], ViT-B/16 [8] and ViT-L/14 [8]. Whereas, CDTrans* has used DeiT-base backbone only. The overall best accuracy and best within per backbone are indicated in bold and box respectively.

Method	f_v	Ar→Cl	Ar→Pr	Ar→Rw	Cl→Ar	Cl→Pr	Cl→Rw	Pr→Ar	Pr→Cl	Pr→Rw	Rw→Ar	Rw→Cl	Rw→Pr	**Avg**
RN-50 [10]	RN-50	34.9	50.0	58.0	37.4	41.9	46.2	38.5	31.2	60.4	53.9	41.2	59.9	46.1
DANN [6]	RN-50	45.6	59.3	70.1	47.0	58.5	60.9	46.1	43.7	68.5	63.2	51.8	76.8	57.6
GSDA [12]	RN-50	61.3	76.1	79.4	65.4	73.3	74.3	65.0	53.2	80.0	72.2	60.6	83.1	70.3
GVB-GD [3]	RN-50	57.0	74.7	79.8	64.6	74.1	74.6	65.2	55.1	81.0	74.6	59.7	84.3	70.4
SPL [38]	RN-50	54.5	77.8	81.9	65.1	78.0	81.1	66.0	53.1	82.8	69.9	55.3	86.0	71.0
SRDC [32]	RN-50	52.3	76.3	81.0	69.5	76.2	78.0	68.7	53.8	81.7	76.3	57.1	85.0	71.3
CLIP [26]	RN-50	51.6	81.9	82.6	71.9	81.9	82.6	71.9	51.6	82.6	71.9	51.6	81.9	72.0
DAPL [7]	RN-50	54.1	84.3	84.8	74.4	83.7	85.0	74.5	54.6	84.8	75.2	54.7	83.8	74.5
AD-CLIP	RN-50	55.4	85.2	85.6	76.1	85.8	86.2	76.7	56.1	85.4	76.8	56.1	85.5	**75.9** $\pm$ 0.1
CDTrans* [39]	ViT-B/16	68.8	85.0	86.9	81.5	87.1	87.3	79.6	63.3	88.2	82.0	66.0	90.6	80.5
TVT [40]	ViT-B/16	74.9	86.8	89.5	82.8	88.0	88.3	79.8	71.9	90.1	85.5	74.6	90.6	83.6
SSRT [31]	ViT-B/16	75.2	89.0	91.1	85.1	88.3	90.0	85.0	74.2	91.3	85.7	78.6	91.-8	85.4
CLIP [26]	ViT-B/16	67.8	89.0	89.8	82.9	89.0	89.8	82.9	67.8	89.8	82.9	67.8	89.0	82.4
DAPL [7]	ViT-B/16	70.6	90.2	91.0	84.9	89.2	90.9	84.8	70.5	90.6	84.8	70.1	90.8	84.0
AD-CLIP	ViT-B/16	70.9	92.5	92.1	85.4	92.4	92.5	86.7	74.3	93.0	86.9	72.6	93.8	**86.1** $\pm$ 0.2
CLIP [26]	ViT-L/14	74.2	93.1	93.3	87.3	93.1	93.3	87.3	74.2	93.3	87.3	74.2	93.1	87.0
DAPL [7]	ViT-L/14	77.3	94.6	94.3	88.6	94.6	94.0	88.8	76.8	94.0	89.0	77.8	94.4	88.7
AD-CLIP	ViT-L/14	**80.3**	**95.4**	**95.7**	**90.9**	**95.5**	**95.2**	**90.1**	**79.6**	**95.1**	**90.8**	**81.1**	**95.9**	**90.5** $\pm$ 0.2

-

Table 2: Comparison of AD-CLIP with state-of-the-art methods for UDA task on VisDA-2017 [25] dataset. We show our results for every class with three different vision backbones. However, CDTrans* has used DeiT-base [33] backbone only. The overall best accuracy and best within per backbone are indicated in bold and box respectively.

Method	f_v	plane	bicycle	bus	car	horse	knife	mcycl	person	plant	sktbrd	train	truck	Avg
RN-101 [10]	RN-101	55.1	53.3	61.9	59.1	80.6	17.9	79.7	31.2	81.0	26.5	73.5	8.5	52.4
DANN [6]	RN-101	81.9	77.7	82.8	44.3	81.2	29.5	65.1	28.6	51.9	54.6	82.8	7.8	57.4
JAN [22]	RN-101	75.7	18.7	82.3	86.3	70.2	56.9	80.5	53.8	92.5	32.2	84.5	54.5	65.7
MODEL [19]	RN-101	94.8	73.4	68.8	74.8	93.1	95.4	88.6	84.7	89.1	84.7	83.5	48.1	81.6
STAR [23]	RN-101	95.0	84.0	84.6	73.0	91.6	91.8	85.9	78.4	94.4	84.7	87.0	42.2	82.7
CLIP [26]	RN-101	98.2	83.9	90.5	73.5	97.2	84.0	95.3	65.7	79.4	89.9	91.8	63.3	84.4
DAPL [7]	RN-101	97.8	83.1	88.8	77.9	97.4	91.5	94.2	79.7	88.6	89.3	92.5	62.0	86.9
AD-CLIP	RN-101	98.1	83.6	91.2	76.6	98.1	93.4	96.0	81.4	86.4	91.5	92.1	64.2	**87.7** $\pm$ 0.2
CDTrans* [39]	ViT-B/16	97.1	90.5	82.4	77.5	96.6	96.1	93.6	**88.6**	**97.9**	86.9	90.3	62.8	88.4
TVT [40]	ViT-B/16	97.1	92.9	85.3	66.4	97.1	97.1	89.3	75.5	95.0	94.7	94.5	55.1	86.7
SSRT [31]	ViT-B/16	98.9	87.6	89.1	**84.8**	98.3	**98.7**	96.3	81.1	94.9	97.9	94.5	43.1	88.8
CLIP [26]	ViT-B/16	99.1	91.7	93.8	76.7	98.4	91.7	95.3	82.7	86.5	96.0	94.6	60.5	88.9
DAPL [7]	ViT-B/16	99.2	92.5	93.3	75.4	98.6	92.8	95.2	82.5	89.3	96.5	95.1	63.5	89.5
AD-CLIP	ViT-B/16	99.6	92.8	94.0	78.6	98.8	95.4	96.8	83.9	91.5	95.8	95.5	65.7	**90.7** $\pm$ 0.3
CLIP [26]	ViT-L/14	99.5	91.1	92.0	69.2	99.2	89.5	97.5	84.3	82.8	98.2	96.9	69.1	89.1
DAPL [7]	ViT-L/14	99.6	91.6	92.9	75.7	99.4	93.3	97.4	84.8	85.5	97.9	97.4	70.5	90.5
AD-CLIP	ViT-L/14	**99.8**	**93.2**	**95.2**	79.1	**99.7**	96.4	**98.5**	86.4	94.0	**98.6**	**98.1**	**73.2**	**92.7** $\pm$ 0.1

of the text embeddings corresponding to the art and clipart domains across the 10 classes of the Office-Home dataset [36]. Figure 4 exhibits the t-SNE visualization of the text embeddings generated by AD-CLIP for the art and clipart domains. The visualization provides an intuitive representation of the distribution and clustering of textual features across different classes and domains.

The t-SNE visualization highlights the simultaneous achievement of domain invariance and discriminativeness by AD-CLIP. The textual embeddings show clear separation between different classes, indicating the discriminative power of the model in distinguishing diverse object categories. Moreover, despite the domain shift between the art and clipart domains, the embeddings demonstrate overlapping regions, signifying the successful domain invariance achieved by AD-CLIP. This capability to maintain similarity between text embeddings from different domains is crucial for effective domain adaptation.

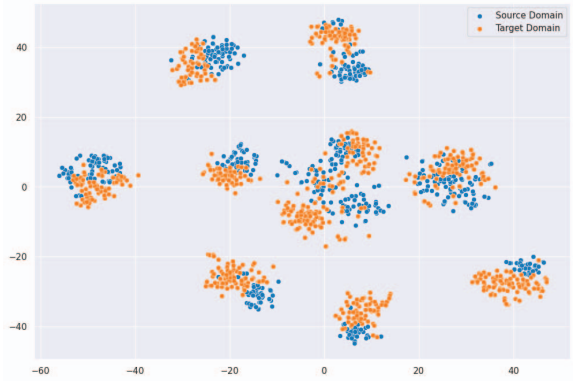

Figure 4: t-SNE visualizations of text embeddings from art and clipart domains of 10 classes of Office-Home.

Sensitivity to the multi-scale features: We also conduct an ablation study to evaluate the sensitivity of AD-CLIP to

visual content and style features obtained from different layers of the CLIP vision backbones. Our aim is to investigate the impact of incorporating multiple f_v layers on the performance of AD-CLIP. Specifically, we utilize three CLIP vision backbones, namely ResNet-50, ViT-B/16, and ViT-L/14, and vary the number of feature layers used to calculate $\hat{\mathcal{F}}(x)$, ranging from the initial to the final layer. To perform the ablation study, we modify the feature extraction process in AD-CLIP by leveraging different f_v layers from the chosen CLIP vision backbones. We begin by extracting visual features from the initial layer and progressively increase the number of layers until reaching the final layer. We then evaluate the impact of these variations on the overall performance of AD-CLIP. Figure 5 presents the results of our ablation study. The plot illustrates the performance trend of AD-CLIP as we incorporate additional layers to extract content features. Remarkably, the results demonstrate a consistent upward trend, indicating that including more layers for content feature extraction leads to improved performance. This observation suggests that the model benefits from incorporating visual information from multiple layers, enabling it to capture more nuanced and discriminative features.

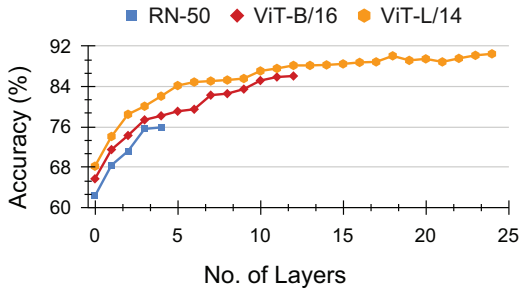

Figure 5: Performance of AD-CLIP with different layers of RN50, ViT-B/16 and ViT-L/14 backbones to extract multi-scale features on Office-Home.

Sensitivity on prompt behaviour and multi-scale feature information: We have analyzed the impact of different prompt settings on the performance of AD-CLIP. The evaluation involved comparing various configurations, including the presence or absence of the domain agnostic token (DAT), the use of manual prompts with image-specific tokens (IST), the source-domain style token (SST) approach, and the full AD-CLIP model with the f_{smn} mechanism. Figure 6 illustrates the results of the ablation study. We observed that omitting the DAT from the prompt led to a decrease in performance, highlighting its importance in guiding AD-CLIP to learn domain-invariant representations. When using manual prompts with IST instead of learned ones, minor improvements were observed, but the overall impact was not significant. Attempting to use an average style information from all source domains as the SST resulted in decreased performance and overfitting of the model. However, our full AD-CLIP model with f_{smn} consistently achieved the best performance across all prompt settings, demonstrating its effectiveness in enhancing domain adaptation and discriminative representation learning. The ablation study provides valuable insights into the prompt configurations of AD-CLIP. The findings underscore the importance of the DAT, dynamic learning of IST and style information through f_{smn}, and highlight the superiority of our proposed AD-CLIP model in achieving robust domain adaptation performance.

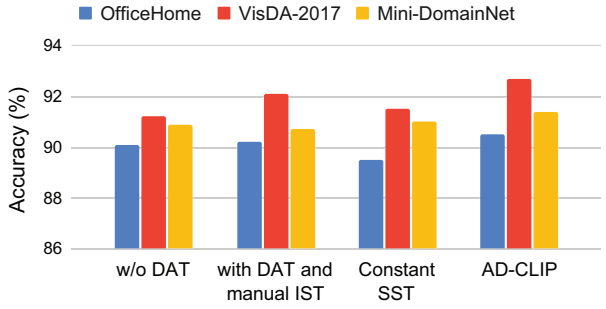

Figure 6: Comparison of results of AD-CLIP with different prompt settings. Here DAT, IST and SST refer to domain-agnostic token, image-specific tokens and source-domain style tokens.

Sensitivity on loss terms: Table 4 presents the results of our ablation study on AD-CLIP, focusing on the influence of various loss terms across all three datasets. The experiments involved omitting the entropy minimization term and the KL divergence loss, both of which led to a significant decrease in performance, emphasizing their crucial role in the optimization process. On the other hand, utilizing all the loss functions consistently boosted the performance of AD-CLIP. Furthermore, we have conducted the experiments under two settings: one involving multi-scale features and the other considering only the features from the final layer of f_v to define content and style information. However, no significant differences in performance were observed between these settings, suggesting that both configurations are equally effective for enhancing AD-CLIP's performance in domain adaptation tasks. The results confirm the significance of the entropy minimization term and KL divergence loss in the loss function of AD-CLIP. The adoption of these loss functions, in conjunction with multi-scale or final layer features, contributes to the model's robustness and demonstrates its potential as a versatile and effective approach for domain adaptation across diverse datasets.

Ablation analysis for Image token length: In order to evaluate the sensitivity of AD-CLIP to the length of image-

Table 3: Comparison of AD-CLIP with the state-of-the-art vision-language models for UDA task on Mini-DomainNet [24] dataset. The overall best accuracy and best within per backbone are indicated in bold and box respectively.

Method	f_v	Cl→Pn	Cl→Rl	Cl→Sk	Pn→Cl	Pn→Rl	Pn→Sk	Rl→Cl	Rl→Pn	Rl→Sk	Sk→Cl	Sk→Pn	Sk→Rl	Avg
CLIP [26]	RN-50	67.9	84.8	62.9	69.1	84.8	62.9	69.2	67.9	62.9	69.1	67.9	84.8	71.2
DAPL [7]		72.4	87.6	65.9	72.7	87.6	65.6	73.2	72.4	66.2	73.8	72.9	87.8	74.8
AD-CLIP		71.7	88.1	66.0	73.2	86.9	65.2	73.6	73.0	68.4	72.3	74.2	89.3	**75.2** $\pm$ 0.2
CLIP [26]	ViT-B/16	80.3	90.5	77.8	82.7	90.5	77.8	82.7	80.3	77.8	82.7	80.3	90.5	82.8
DAPL [7]		83.3	92.4	81.1	86.4	92.1	81.0	86.7	83.3	80.8	86.8	83.5	91.9	85.8
AD-CLIP		84.3	93.7	82.4	87.5	93.5	82.4	87.3	84.5	81.6	87.9	84.8	93.0	**86.9** $\pm$ 0.2
CLIP [26]	ViT-L/14	85.2	92.4	86.2	89.2	92.4	86.2	89.2	85.2	86.2	89.2	85.2	92.4	88.3
DAPL [7]		86.8	93.5	87.9	90.5	93.5	88.3	90.2	87.8	88.6	90.0	86.8	93.5	89.8
AD-CLIP		**89.1**	**94.5**	**89.2**	**91.9**	**95.0**	**90.1**	**92.0**	**89.2**	**90.3**	**92.3**	**88.4**	**95.1**	**91.4** $\pm$ 0.1

specific context tokens $V_{l\,l=1}^{L}$ obtained from C_v, we conducted experiments on three benchmark datasets for unsupervised domain adaptation (UDA) tasks. Figure 7 displays the results of varying the length of image tokens.

The findings of the ablation study reveal that the best performance is achieved at an optimal length of $L = 4$ for the image-specific context tokens. This indicates that AD-CLIP benefits from considering a moderate number of image tokens to effectively capture relevant visual information and enhance its performance in domain adaptation tasks across diverse datasets.

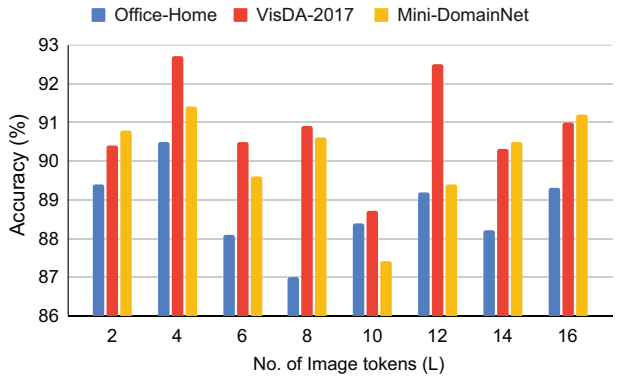

Figure 7: Performance of AD-CLIP with different numbers of image tokens in the prompt

Table 4: Ablation study of AD-CLIP with different losses in three datasets using source encoder ViT-L/14 and source-assisted encoder ViT-B/16. Here 'w-ms' defines ablation with multi-scale features and 'w/o-ms' defines without multi-scale features.

Loss	Office-Home		VisDA-2017		Mini-DomainNet	
	w-ms	w/o-ms	w-ms	w/o-ms	w-ms	w/o-ms
$\mathbf{L}_{ce}$ (no adaptation)	87.6	87.4	89.4	89.3	88.7	88.6
$\mathbf{L}_{ce} + \mathbf{L}_{smn}$ (no adaptation)	87.9	87.2	90.1	89.5	89.3	89.2
$\mathbf{L}_{ce} + \mathbf{L}_{smn} + \mathbf{L}_{em}$	88.1	87.7	89.8	89.5	89.6	89.4
$\mathbf{L}_{ce} + \mathbf{L}_{Align}$	89.1	89.2	91.0	90.9	90.6	90.8
$\mathbf{L}_{ce} + \mathbf{L}_{smn} + \mathbf{L}_{Align}$	**90.5**	90.1	**92.7**	91.9	**91.4**	91.1

Model Complexity We train our model on NVIDIA RTX A6000 GPU with 48 GB card. Results indicate that AD-CLIP requires 17.2%, 0.54%, 0.18% more computational resources than DANN[6], CLIP[26] and DAPL[7] respectively as shown in Figure 8. However, AD-CLIP outperforms every state-of-the-art method in UDA task.

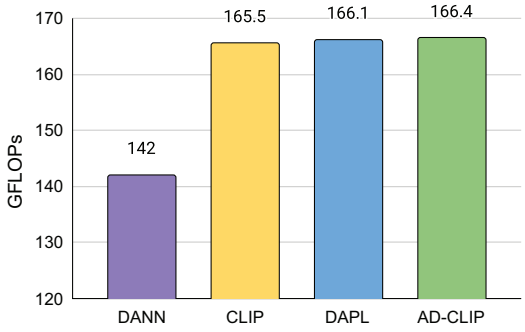

Figure 8: Comparison of the computational complexity of AD-CLIP with other UDA methods in terms of GFLOPs.

5. Takeaways

In this paper, we propose a novel framework called AD-CLIP that tackles the unsupervised DA problem through prompt learning for foundation models. Our approach is based on the CLIP model and focuses on learning domain-invariant and class-generic prompt tokens using visual space features. To achieve this, we leverage the vision encoder of CLIP to extract multi-scale style and content features and adapt them to target datasets using learnable projector networks. Specifically, we learn three types of tokens in the prompts per image: domain token, image token, and class token. Additionally, we introduce a combination of distribution divergence loss and entropy minimization loss to align domains. Our experimental results on three benchmark DA datasets demonstrate that AD-CLIP outperforms existing state-of-the-art methods. In the future, we plan to extend our approach to solve specific applications such as person re-identification and medical imaging.

References

[1] Mahsa Baktashmotlagh, Mehrtash T Harandi, Brian C Lovell, and Mathieu Salzmann. Unsupervised domain adaptation by domain invariant projection. In *Proceedings of the IEEE international conference on computer vision*, pages 769–776, 2013.

[2] Shirsha Bose, Enrico Fini, Ankit Jha, Mainak Singha, Biplab Banerjee, and Elisa Ricci. Stylip: Multi-scale style-conditioned prompt learning for clip-based domain generalization, 2023.

[3] Shuhao Cui, Shuhui Wang, Junbao Zhuo, Chi Su, Qingming Huang, and Qi Tian. Gradually vanishing bridge for adversarial domain adaptation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12455–12464, 2020.

[4] Hal Daumé III, Abhishek Kumar, and Avishek Saha. Frustratingly easy semi-supervised domain adaptation. In *Proceedings of the 2010 Workshop on Domain Adaptation for Natural Language Processing*, pages 53–59, 2010.

[5] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[6] Yaroslav Ganin and Victor Lempitsky. Unsupervised domain adaptation by backpropagation. In *International conference on machine learning*, pages 1180–1189. PMLR, 2015.

[7] Chunjiang Ge, Rui Huang, Mixue Xie, Zihang Lai, Shiji Song, Shuang Li, and Gao Huang. Domain adaptation via prompt learning. *arXiv preprint arXiv:2202.06687*, 2022.

[8] Kai Han, Yunhe Wang, Hanting Chen, Xinghao Chen, Jianyuan Guo, Zhenhua Liu, Yehui Tang, An Xiao, Chunjing Xu, Yixing Xu, et al. A survey on vision transformer. *IEEE transactions on pattern analysis and machine intelligence*, 2022.

[9] David Haussler and Manfred Warmuth. The probably approximately correct (pac) and other learning models. *Foundations of knowledge acquisition: Machine learning*, pages 291–312, 1993.

[10] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[11] Judy Hoffman, Eric Tzeng, Taesung Park, Jun-Yan Zhu, Phillip Isola, Kate Saenko, Alexei A. Efros, and Trevor Darrell. Cycada: Cycle-consistent adversarial domain adaptation, 2017.

[12] Lanqing Hu, Meina Kan, Shiguang Shan, and Xilin Chen. Unsupervised domain adaptation with hierarchical gradient synchronization. In *Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition*, pages 4043–4052, 2020.

[13] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International Conference on Machine Learning*, pages 4904–4916. PMLR, 2021.

[14] Muhammad Uzair Khattak, Hanoona Rasheed, Muhammad Maaz, Salman Khan, and Fahad Shahbaz Khan. Maple: Multi-modal prompt learning. *arXiv preprint arXiv:2210.03117*, 2022.

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

[16] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Communications of the ACM*, 60(6):84–90, 2017.

[17] Jingjing Li, Erpeng Chen, Zhengming Ding, Lei Zhu, Ke Lu, and Heng Tao Shen. Maximum density divergence for domain adaptation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(11):3918–3930, 2021.

[18] Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. Visualbert: A simple and performant baseline for vision and language. *arXiv preprint arXiv:1908.03557*, 2019.

[19] Rui Li, Qianfen Jiao, Wenming Cao, Hau-San Wong, and Si Wu. Model adaptation: Unsupervised domain adaptation without source data. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9641–9650, 2020.

[20] Yanghao Li, Naiyan Wang, Jiaying Liu, and Xiaodi Hou. Demystifying neural style transfer. *arXiv preprint arXiv:1701.01036*, 2017.

[21] Mingsheng Long, Zhangjie Cao, Jianmin Wang, and Michael I Jordan. Conditional adversarial domain adaptation. *Advances in neural information processing systems*, 31, 2018.

[22] Mingsheng Long, Han Zhu, Jianmin Wang, and Michael I. Jordan. Deep transfer learning with joint adaptation networks, 2017.

[23] Zhihe Lu, Yongxin Yang, Xiatian Zhu, Cong Liu, Yi-Zhe Song, and Tao Xiang. Stochastic classifiers for unsupervised domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9111–9120, 2020.

[24] Xingchao Peng, Qinxun Bai, Xide Xia, Zijun Huang, Kate Saenko, and Bo Wang. Moment matching for multi-source domain adaptation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1406–1415, 2019.

[25] Xingchao Peng, Ben Usman, Neela Kaushik, Judy Hoffman, Dequan Wang, and Kate Saenko. Visda: The visual domain adaptation challenge. *arXiv preprint arXiv:1710.06924*, 2017.

[26] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021.

[27] Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. Improving language understanding by generative pre-training. 2018.

[28] Kuniaki Saito, Kohei Watanabe, Yoshitaka Ushiku, and Tatsuya Harada. Maximum classifier discrepancy for unsupervised domain adaptation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3723–3732, 2018.

[29] Taylor Shin, Yasaman Razeghi, Robert L Logan IV, Eric Wallace, and Sameer Singh. Autoprompt: Eliciting knowledge from language models with automatically generated prompts. *arXiv preprint arXiv:2010.15980*, 2020.

[30] Mainak Singha, Ankit Jha, Bhupendra Solanki, Shirsha Bose, and Biplab Banerjee. Applenet: Visual attention parameterized prompt learning for few-shot remote sensing image generalization using clip. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 2023–2033, June 2023.

[31] Tao Sun, Cheng Lu, Tianshuo Zhang, and Haibin Ling. Safe self-refinement for transformer-based domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7191–7200, 2022.

[32] Hui Tang, Ke Chen, and Kui Jia. Unsupervised domain adaptation via structurally regularized deep clustering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8725–8735, 2020.

[33] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention, 2021.

[34] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008.

[35] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[36] Hemanth Venkateswara, Jose Eusebio, Shayok Chakraborty, and Sethuraman Panchanathan. Deep hashing network for unsupervised domain adaptation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5018–5027, 2017.

[37] Mei Wang and Weihong Deng. Deep visual domain adaptation: A survey. *Neurocomputing*, 312:135–153, 2018.

[38] Qian Wang and Toby Breckon. Unsupervised domain adaptation via structured prediction based selective pseudo-labeling. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 6243–6250, 2020.

[39] Tongkun Xu, Weihua Chen, Pichao Wang, Fan Wang, Hao Li, and Rong Jin. Cdtrans: Cross-domain transformer for unsupervised domain adaptation. *arXiv preprint arXiv:2109.06165*, 2021.

[40] Jinyu Yang, Jingjing Liu, Ning Xu, and Junzhou Huang. Tvt: Transferable vision transformer for unsupervised domain adaptation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 520–530, 2023.

[41] Ting Yao, Yingwei Pan, Chong-Wah Ngo, Houqiang Li, and Tao Mei. Semi-supervised domain adaptation with subspace learning for visual recognition. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 2142–2150, 2015.

[42] Sicheng Zhao, Xiangyu Yue, Shanghang Zhang, Bo Li, Han Zhao, Bichen Wu, Ravi Krishna, Joseph E Gonzalez, Alberto L Sangiovanni-Vincentelli, Sanjit A Seshia, et al. A review of single-source deep unsupervised visual domain adaptation. *IEEE Transactions on Neural Networks and Learning Systems*, 33(2):473–493, 2020.

[43] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Conditional prompt learning for vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16816–16825, 2022.

[44] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *International Journal of Computer Vision*, 130(9):2337–2348, 2022.