Domain Adaptation via Prompt Learning

Chunjiang Ge, Rui Huang, Mixue Xie, Zihang Lai, Shiji Song, Senior Member, IEEE, Shuang Li, and Gao Huang, Member, IEEE

Abstract—Unsupervised domain adaptation (UDA) aims to adapt models learned from a well-annotated source domain to a target domain, where only unlabeled samples are given. Current UDA approaches learn domain-invariant features by aligning source and target feature spaces through statistical discrepancy minimization or adversarial training. However, these constraints could lead to the distortion of semantic feature structures and loss of class discriminability. In this article, we introduce a novel prompt learning paradigm for UDA, named domain adaptation via prompt learning (DAPrompt). In contrast to prior works, our approach learns the underlying label distribution for target domain rather than aligning domains. The main idea is to embed domain information into prompts, a form of representation generated from natural language, which is then used to perform classification. This domain information is shared only by images from the same domain, thereby dynamically adapting the classifier according to each domain. By adopting this paradigm, we show that our model not only outperforms previous methods on several cross-domain benchmarks but also is very efficient to train and easy to implement.

Index Terms—Contrastive learning, prompt learning, unsupervised domain adaptation (UDA).

I. INTRODUCTION

Deep learning has achieved great success in recent years [1], [2], [3] with the help of large-scale annotated datasets [4]. Since annotating large-scale datasets is costly and time-consuming, researchers propose to train a model for an unlabeled domain by leveraging a related domain that is well-annotated. However, a model (e.g., a neural network) trained on an annotated domain may not generalize well to an unlabeled domain due to distribution shift [5], [6], [7]. The problem of unsupervised domain adaptation (UDA) [8], [9], [10], [11] has been proposed to study the transferring of knowledge under such domain shift.

Conventional UDA methods mainly resort to learning domain-invariant representations by minimizing statistical discrepancy [12], [13], [14], [15] or applying adversarial training [16], [17], [18], [19], [20] to minimize the distribution shift between the source and target domains. With similar feature distribution led by domain alignment, the classifier trained on the source domain can be directly applied to the target data (see the top of Fig. 1). However, reducing the discrepancy by aligning domains could lead to a loss of semantic information [21], [22]. For example, $f(X_T)$ would not contain any color information since $f(X_T)$ is aligned with grayscale image distribution $f(X_S)$ (see the top of Fig. 1). Furthermore, such loss of semantic information could harm class discriminability since discarded domain-specific knowledge is useful in classification [23], [24]. Even some recent UDA methods [21], [24], [25], [26] advocate preserving the semantic information to maintain the class discriminability. However, these methods suffer from a subtle tradeoff between domain alignment and preserving semantic features [22], [23], [27] as two objectives could be adversarial.

To remedy this, we propose domain adaptation via prompt learning (DAPrompt), which preserves domain-specific knowledge and embeds it into prompt representation (“sketch” and “dog” in Fig. 1, bottom) to avoid domain alignment. Specifically, the prompt consists of three parts: domain-agnostic context, domain-specific context, and class label (token). Fig. 2(a) shows our prompt design. Each image corresponds to a ground-truth class through the domain and class information of prompt. For example, an image that shows “an art work of a dog” could correspond to the prompt “an image of a painting dog.” The domain-agnostic context represents general task information and is shared among all images. The domain-specific context represents domain information and is shared within each domain. The class label distinguishes different categories. Our domain-specific prompt representation learns the conditional probability distribution of source and target domains, respectively. Furthermore, it allows for a dynamic mechanism to adapt the classifier to the source or target domain. More details are discussed in Section III-B.

To learn domain and category disentangled prompt representation, we apply a contrastive objective for training. An image and a prompt form a pair of positive example only when the domain and category of them are matched, respectively, while any other cases are negative examples. By contrasting the representation of $X_S$ and $y$, the representation of the image and prompt are aligned in the feature space, respectively (Fig. 1, bottom). Furthermore, the prompt representation of “sketch” is pushed away from the “photo” domain by contrasting $X_T$ and $y$. More details are discussed in Section III-D.

To enable contrastive learning in the target domain,
that our method consistently yields promising performance. As shown in Fig. 2(b), we empirically find that fine-tuning and linear probing of CLIP both fail to generalize well on the target domain as their performance is lower than zero-shot CLIP. Prompt learning performs better but suffers from distribution shift. Our DAPrompt outperforms all the baseline models since our domain-specific prompt improves the performance. On VisDA2017, our method has 2.5%/1.1% and 4.9%/4.1% absolute improvement with respect to CLIP and CoOp (visual encoder is ResNet101/ViT-B/16). On miniDomainNet and Office-Home, DAPrompt achieves an overall accuracy of 74.8%/85.8% and 74.5%/84.4% (ResNet50/ViT-B/16), which are 3.6%/3.0% and 2.5%/2.0% points higher than zero-shot CLIP model. We also achieve an sota performance of 74.5%/86.9% on Office-Home [28] and VisDA-2017 [29]. To summarize, the contributions of our work are threefold.

1) We propose DAPrompt for unsupervised domain adaptation. It is simple yet effective, easy to implement, and fast for training. Our method builds a bridge between the multimodal method and UDA. To the best of our knowledge, we are the first to explore the application of multimodal method and prompt learning in UDA.

2) We propose to use domain-specific prompt for UDA and devise a dynamic mechanism to adapt the classifier to each domain. Our method aims at learning the underlying label representation of target domain rather than aligning domains. Contrastive objectives ensure the disentangled representations for each category and domain.

3) Our DAPrompt has significantly improved the performance on VisDA2017, miniDomainNet, and Office-Home with respect to CLIP and CoOp. It has achieved the state-of-the-art performance on the Office-Home and VisDA-2017 datasets, improving the accuracy by 3.2%/4.2% over the sota methods and 2.5%/2.5% over CLIP.

II. RELATED WORK

A. Unsupervised Domain Adaptation

UDA adapts a model trained on a labeled source domain to an unlabeled target domain [30], [31], [32]. Quite a few UDA methods learn domain-invariant features via minimizing the discrepancy between domains [13], [15], [33], [34], [35]. For example, Tzeng et al. [12] introduced an adaptation layer and a domain confusion loss to learn semantically meaningful and domain-invariant representations. DAN [13] aligns source and target domains by minimizing the maximum mean discrepancy (MMD) on task-specific layers. Sun and Saenko [33] proposed CORAL that aligns the second-order statistics of the source and target domains with a linear projection.

Inspired by generative adversarial networks (GANs) [16], another family of UDA methods applies adversarial learning to obtain domain-invariant representations [9], [17], [36], [37]. For example, DANN [9] and CDAN [17] introduce a domain discriminator to distinguish source samples from target ones, while the feature extractor tries to generate domain-invariant features in order to fool the domain discriminator. Differently,
MCD [38] plays the minimax game between a feature encoder and two classifiers, where two classifiers try to maximize their prediction discrepancy and the feature extractor aims to minimize that discrepancy.

Despite the success achieved by domain alignment, class discrimination also loses due to the distorted structure of semantic features [21], [23]. How to maintain class discrimination also loses due to the distorted structure of their prediction discrepancy and the feature extractor aims to and two classifiers, where two classifiers try to maximize MCD [38] plays the minimax game between a feature encoder and two classifiers, where two classifiers try to maximize their prediction discrepancy and the feature extractor aims to minimize that discrepancy.

Despite the success achieved by domain alignment, class discrimination also loses due to the distorted structure of semantic features [21], [23]. How to maintain class discrimination also loses due to the distorted structure of their prediction discrepancy and the feature extractor aims to minimize that discrepancy.

To name a few, Chen inability has also been considered by recent UDA works [21], [23]. How to maintain class discrimination also loses due to the distorted structure of their prediction discrepancy and the feature extractor aims to minimize that discrepancy.

Our work proposes to learn domain-specific prompt disentanglement of semantic and domain features attained by our DAPrompt and a dynamic mechanism (Section III-B), and pseudo labeling (Section III-C) and discuss the disentanglement of semantic and domain features attained by our DAPrompt method (Section III-D).

### A. Preliminaries

An overview of our method is shown in Fig. 3. Our model is composed of an image encoder $f(\cdot)$ and a text encoder $g(\cdot)$. The image encoder can be a ResNet [1] or Vision Transformer (ViT) [52], and the text encoder is a Transformer [53]. The image and text input can be directly transformed from a high-dimensional space into a low-dimensional feature space by the encoders.

Our method is trained with image–text pairs in a contrastive manner. For example, an input text describes a category in the format of “a photo of a [CLASS]” ([CLASS] is the class token). A positive pair is an image $x$ with its corresponding
text \( t_i \) describing the category of \( x_i \). A negative pair is an image \( x_i \) with an irrelevant description \( t_j, j \neq i \) in the mini-batch. The training objective is to maximize the cosine similarity of positive pairs and minimize the cosine similarity of negative pairs. The contrastive learning objective aligns the image and text representation in the same feature space.

With the aligned features, the model is capable of performing zero-shot inference. By forwarding \( K \) category descriptions, an image \( x \) would belong to the category \( \hat{y}_i \) with the largest similarity

\[
P(\hat{y} = i | x, t) = \frac{\exp(\langle g(t_i), f(x) \rangle/T)}{\sum_{k=1}^{K} \exp(\langle g(t_k), f(x) \rangle/T)} \tag{1}
\]

\[
\hat{y}_i = \text{arg max } P(\hat{y} = k) \tag{2}
\]

where \( T \) is a user-defined hyperparameter (temperature) and \( \langle \cdot, \cdot \rangle \) denotes the cosine similarity.

The input text described above is a manually designed prompt comprised of a sequence of discrete tokens. The manually designed prompts are transformed into fixed vectors in the word embedding space. Since these word embeddings could be suboptimal for this specific task, optimizing the continuous embedding of the tokens has been proposed [44], [51]. The continuous representation \( t_k \) allows for a more precise description of semantic features that are important for class discrimination.

Existing prompt learning methods adopt a domain-agnostic style that context is shared across all domains and all categories. It has a unified format

\[
t_k = [v_1]_1[v_2]_1 \cdots [v_{M_1}]_1[CLASS]_k \tag{3}
\]

where \([v_m]_i, m_1 \in [1, 2, \ldots, M_1]\), is a vector with the same dimension as the word embedding and \( M_1 \) is the number of context tokens applied in the prompt.

**B. Learning With Domain-Specific Prompt**

In this section, we describe the details on domain-specific prompt and the dynamic mechanism to adapt classifier to source or target domain. The overview of our method is shown in Fig. 3. Since we notice that the class labels for source and target domains could be different semantically, the domain-agnostic context ignores the discrepancy between domains. To remedy this, we propose to design domain-specific prompt in the prompt to deal with distribution shift [shown in Fig. 3(a)]. To be specific, our proposed prompt contains three parts, domain-agnostic tokens, domain-specific tokens, and class tokens. We use \([d]^{d_1}_{m_1}, m_2 \in [1, 2, \ldots, M_2]\), to denote domain-specific tokens, which have the same dimension as word embeddings. The domain-specific tokens are shared among all categories but specially designed for each domain \([d]^{d_i}_j \neq [d]^{d_j}_i, i, j \in [1, 2, \ldots, M_2]\). The number of domain-specific tokens is denoted by \( M_2 \). Domain indicator denotes the source and target domains \( d \in \{s, t\} \). The overall prompt is defined in the following format:

\[
t^{d_k}_i = [v]_1[v]_2 \cdots [v]_{M_1}[d]^{d_1}_{m_1}[d]^{d_2}_{m_2} \cdots [d]^{d_{M_2}}_{M_2}[CLASS]_k. \tag{4}
\]

Furthermore, a trainable class-aware prompt could learn fine-grained category representation. The domain-agnostic context could follow a class-specific style denoted by class-specific context. Each class could be initialized with different tokens

\[
t^{d_k}_i = [v]_1[v]_2 \cdots [v]_{M_1}[d]^{d_1}_{m_1}[d]^{d_2}_{m_2} \cdots [d]^{d_{M_2}}_{M_2}[CLASS]_k. \tag{5}
\]
As we mentioned before, an image and prompt form a positive pair only when category and domain match, respectively. Hence, the domain-specific tokens are updated with positive pairs from the corresponding domain. Therefore, domain-specific tokens embed the information shared within each domain. It is worth noting that even though we apply “text” \( [t^u_i] \) for classification, our method does not require any extra information. Trainable domain-agnostic \( [v^u_i] \) and domain-specific context \( [d^u_i], [d^s_i] \) for source and target domains are randomly initialized. Class tokens [CLASS] are provided by the class names of the dataset.

We have 2\( K \) categories in total since we apply different prompts \( t^s_k \) and \( t^u_k \) for the source and the target domain, respectively. Given a set of training samples \( \{x^s_i, y^s_i\}_{i=1}^{N^s} \) and unlabeled \( \{x^u_i\}_{i=1}^{N^u} \), we could obtain the probability that a sample belongs to the \( k \)th category according to

\[
P(\hat{y}^s_i = k|x^s_i, t^s_k) = \frac{\exp((g(t^s_k), f(x^s_i))/T)}{\sum_{d \in \{s,u\}} \sum_{j=1}^{K} \exp((g(t^s_j), f(x^s_i))/T)}
\]

(6)

\[
P(\hat{y}^u_i = k|x^u_i, t^u_k) = \frac{\exp((g(t^u_k), f(x^u_i))/T)}{\sum_{d \in \{s,u\}} \sum_{j=1}^{K} \exp((g(t^u_j), f(x^u_i))/T)}
\]

(7)

Since we know that images are from source domain or target domain, we introduce a dynamic mechanism to adapt the way of attaining the probability that an image belongs to a category, i.e., \( P(\hat{y}^s_i = k|x^s_i, t^s_k) \) for source domain images and \( P(\hat{y}^u_i = k|x^u_i, t^u_k) \) for target domain images. With the probability of the image \( x_i \) belonging to class \( k \), we minimize the cross-entropy loss given the ground-truth label \( y^s_i \). The cross-entropy loss for the source domain is shown in (8) and we would introduce the loss for target domain in Section III-C

\[
L_s = -\frac{1}{N^s} \sum_{i=1}^{N^s} \log P(\hat{y}^s_i = y^s_i).
\]

(8)

Existing domain adaptation methods train their classifier on the source domain to learn a conditional probability distribution \( P(y|x^s, t^s) \). By aligning the marginal distribution of \( P(f(x^s)) \) and \( P(f(x^u)) \), they could directly make use of the conditional probability to inference on the target domain. These methods suffer from loss of class discriminability [24], [25]. Our method does not align marginal distributions but learns two conditional probability distributions \( P(y|x^s, t^s) \) and \( P(y|x^u, t^u) \) by learning two sets of prompts \( t^s_k, t^u_k, k \in \{1, 2, \ldots, K\} \), respectively. Hence, our method would not suffer from loss of semantic information and could make full use of the information of data.

We freeze the image encoder in our method according to the setting of CoOp. However, freezing weights may limit the capacity of the proposed method. Fine-tuning the image encoder with proper design could further promote the performance. Mask image modeling is a useful pretraining method for image representation and combining it with our method could further improve the quality of representation learning. We leave these as future work.

Algorithm 1 Algorithm of DAPrompt

\begin{itemize}
  \item \textbf{Input:} Vision encoder \( f() \), text encoder \( g() \), source domain data \( x^s \), target domain data \( x^u \), source domain prompt \( t^s \), target domain prompt \( t^u \), label for source domain data \( y^s \);
  \item \textbf{Output:} target domain prompt \( t^u \)
\end{itemize}

1: while not converge do
2: \hspace{1em} Predict source domain image \( x^s_i \) labels by Eq. 6 and \( \hat{y}^s_i = \arg \max_k P(\hat{y}^s_i = k|x^s_i, t^s_k) \);
3: \hspace{1em} Predict target domain image \( x^u_i \) labels by Eq. 7 and \( \hat{y}^u_i = \arg \max_k P(\hat{y}^u_i = k|x^u_i, t^u_k) \);
4: \hspace{1em} Predict pseudo label for target domain image \( x^u_i \) with zero-shot inference by CLIP, \( \hat{y}^u_i = \arg \max_k P(\hat{y}^u_i = k|x^u_i) \);
5: \hspace{1em} Calculate the loss of by: \( L = -\log P(\hat{y}^u_i = y^u_i) - \log P(\hat{y}^u_i = \hat{y}^u_i) \) and update source domain prompt \( t^s \), target domain prompt \( t^u \)
6: end while

C. Pseudo Labeling

To further exploit the unlabeled data, we generate pseudo labels on the target domain. We choose from \( K \) classes with maximum predicted probability as the pseudo label \( y^u \) of the training data \( x^u \)

\[
y^u = \arg \max_k P(\hat{y}^u = k|x^u), \quad k = \{1, 2, \ldots, K\}.
\]

(9)

We only generate pseudo labels for unlabeled data whose maximum prediction probability is larger than a fixed threshold \( \tau \) for the quality of pseudo labels. We make use of the zero-shot inference ability of CLIP to generate pseudo labels as described in Section III-A. We train the prompt of target domain \( t^u_k \) with these unlabeled images and their pseudo labels with the contrastive objective (7)

\[
L_u = \frac{1}{N_u} \sum_{i=1}^{N_u} \{P(\hat{y}^u_i = y^u_i) \geq \tau \} \log P(\hat{y}^u_i = y^u_i|x^u_i)
\]

(10)

where \( \{\cdot\} \) is an indicator function. Overall, our proposed DAPrompt method could be trained in an end-to-end manner with a total contrastive loss

\[
L = L_s(D^s) + L_u(D^u).
\]

(11)

We summarize our proposed DAPrompt method in Algorithm 1. We first predict the labels of the source domain data with (6). Then, we predict the labels of the target domain data with (7). We generate pseudo labels for the unlabeled data and train the model with the contrastive loss. We repeat the above steps until convergence.

D. Disentanglement by Contrastive Learning

In this section, we provide an intuitive explanation for why contrastive learning achieves disentanglement in semantic and domain information. First, we assume that the visual representation \( f(x^u_i) \) contains two parts: domain information of domain \( d \) and the intrinsic class information of class \( c \) (Fig. 4(a), \( z_d \) and \( z_c \)). Similarly, the language embedding \( g(t^u_k) \)
A. Datasets and Experimental Settings

**VisDA-2017** [29] is a challenging dataset for synthetic-to-real domain adaptation with 12 categories. It contains 152,397 synthetic images, generated by rendering the 3-D models with different angles and light conditions, and 55,388 real-world images, collected from MSCOCO [54]. Following [17] and [38], we use the synthetic images as the source domain and real-world images as the target domain. **miniDomainNet** [55], [56] is a subset of DomainNet, which is a large-scale cross-domain dataset. It contains 18,703 images of clip art domain (C), 31,202 images of painting domain (P), 65,609 images of real world domain (R), and 24,492 images of sketch domain (S). Each domain contains 126 categories, which has a larger diversity than the above two datasets. The 12 UDA tasks are marked by: i.e., R→S, . . . , P→R.

**Office-Home** [28] is a large-scale benchmark for visual cross-domain recognition. It collects a total of 15,500 images from four different domains: art (A), clip art (C), product (P), and real world (R). Besides, each domain contains the objects of 65 categories in the office and home environments. To evaluate our method, we conduct 12 UDA tasks on the Office-Home dataset, i.e., A→C, . . . , R→P.

B. Implementation Details

Since our method is generic to any image encoders, we choose a CNN and ViT encoder for each dataset. For **VisDA-2017** [29], the results are obtained by leveraging the pretrained CLIP model with ResNet-101 [1] and ViT-B/16 [52] as the image encoder. The parameters of the image and text encoders are fixed and we train the prompt for 25 epochs using the mini-batch stochastic gradient descent (SGD) optimizer with a batch of 32. The learning rate is set to 0.003 initially and decayed with a cosine annealing rule. For **miniDomainNet** [56], we apply the CLIP model with ResNet-50 and ViT-B/16 [52] as its image encoder. We train the model for 15 epochs and use a mini-batch size of 32 with an SGD optimizer. We adopt 0.003 as the learning rate and it decays according to a cosine rule. For **Office-Home**, we use pretrained CLIP model and adopt ResNet-50 [1] as its image encoder. We fix the parameters in the encoders and the prompt is trained with the mini-batch SGD optimizer for 200 epochs, where the batch size is set to be 32. The initial learning rate is set to 0.003 and decayed with a cosine annealing rule [57]. As for the hyperparameters, the lengths of context tokens $M_1$ and domain-specific tokens $M_2$ are both set to 16. Other choices of token numbers are discussed in Section IV-C. Our context vectors are randomly initialized using a zero-mean Gaussian distribution with a standard deviation of 0.02. The pseudo labeling threshold $\tau$ is set to 0.5 for **VisDA-2017** [29], 0.7 for miniDomainNet [56], and 0.6 for Office-Home [28]. Further discussion about the value of $\tau$ is shown in Section IV-C.

C. Ablation Study

To give a detailed analysis of our method, we conduct detailed ablation studies on **VisDA-2017** [29]. All of the variant models are trained with the same training hyperparameters as described in Section IV-B.

---

GE et al.: DOMAIN ADAPTATION VIA PROMPT LEARNING

IV. EXPERIMENTAL RESULTS

We conduct extensive experiments on UDA benchmarks to verify the validity of our proposed method. We next present the datasets used in our experiments, implementation details, ablation studies of our method, comparisons with current methods, and visualization of results.
TABLE I
ABSTRACTION OF OUR PROPOSED DAPROMPT ON THE VISTA-2017 DATASET. HIGHER VALUES ARE BETTER. OUR METHOD IS HIGHLIGHTED WITH GRAY. (a) ABSTRACTION ON DOMAINT-SPECIFIC CONTEXT AND PSEUDO LABELING. WE ADOPT RESNET101 AND ViTB/B16 AS OUR IMAGE ENCODER. FT: FINE-TUNING. LP: LINEAR PROBING. DS: DOMAINT-SPECIFIC PROMPT. (b) ABSTRACTION ON DIFFERENT COMBINATIONS OF TOKEN LENGTH (M1 AND M2) WITH RESNET101. (c) ABSTRACTION ON DIFFERENT COMBINATIONS OF THRESHOLD (M1, M2) WITH RESNET101

| Method | Prompt Template | Domain-agnostic | Domain-specific | Pseudo labeling | Acc. ResNet101 | Acc. ViT-B/16 |
|--------|-----------------|-----------------|----------------|-----------------|---------------|--------------|
| CLIP [3] | Manual | ✔️ | ✔️ | 84.4 | 88.7 |
| +LP | Manual | ✔️ | ✔️ | 74.0 | 80.6 |
| +FT | Manual | ✔️ | ✔️ | 74.5 | 80.5 |
| CoOp [51] | Unified | ✔️ | ✔️ | 82.0 | 85.7 |
| | Class-specific | ✔️ | ✔️ | 82.0 | 85.6 |
| DAPrompt | Unified | ✔️ | ✔️ | 83.0 | 88.3 |
| -DS | Class-specific | ✔️ | ✔️ | 85.6 | 88.7 |
| DAPrompt | Unified | ✔️ | ✔️ | 86.9 | 89.6 |
| | Class-specific | ✔️ | ✔️ | 86.9 | 89.8 |

(a)

1) Ablation on DAPrompt and Other CLIP Variants:
To prove the effectiveness and necessity of domain-specific context and pseudo labeling, we compare the performances of these following methods on the VisDA-2017 [29] dataset.

1) CLIP and Its Variants: We adopt zero-shot CLIP as a baseline model. Since zero-shot CLIP ignores the source data, we conduct linear probing and fine-tuning of CLIP with pseudo labeling to make full use of the dataset. Pseudo labels are generated by zero-shot CLIP with manually designed prompt “a photo of [CLASS],” which is the same with our DAPrompt. The loss is also the same with DAPrompt [see (11)].

2) CoOp: It applies prompt learning with only domain-agnostic prompt. The learnable prompt could be either in unified style [see (3)] or class-specific style. Zero-shot CLIP and CoOp ignore data in the target domain. Hence, these methods both suffer from distribution shift.

3) DAPrompt Without Domain-Specific Prompt (DAPrompt-DS): We remove the domain-specific prompt from DAPrompt and it could also be seen as CoOp with pseudo labeling. In this setting, domain information would be learned entangled with category.

4) DAPrompt With Domain-Specific Prompt [See (5)]: With domain-specific prompt, the model is now able to dynamically adapt the classifier according to (6) and (7). Since removing pseudo labeling method would leave domain-specific prompt trained with only negative examples, we do not conduct such ablation.

The results of the above experiments are listed in Table I(a).

We have the following conclusions.

1) Zero-shot CLIP is very strong baseline. Zero-shot CLIP achieves 84.4%/88.7% accuracy on ResNet101 and ViT-B/16 visual encoders and outperforms nearly all other variants other than DAPrompt.

2) Simple fine-tuning and linear probing could not make use of the pretraining of CLIP. Fine-tuning may distort the learned features of CLIP and performs poorly than zero-shot CLIP. Linear probing also performs poorly for its limited capacity (10.4%/8.1% lower than zero-shot CLIP). We observe that both fine-tuning and linear probing are not the correct way to make use of CLIP in UDA.

3) CoOp suffers from distribution shift. CoOp learns a set of meaningful prompts for the source domain. We observe that CoOp outperforms fine-tuning and linear probing of CLIP on both ResNet101 and ViT-B/16. However, it has to directly apply these prompts on the target domain and suffers from distribution shift, which results in mostly lower performance than zero-shot CLIP (82.0%/85.7% versus 84.4%/88.7%).

4) Pseudo labeling enables the model to fit the data of the target domain. We observe that DAPrompt-DS performs better than CoOp for pseudo labeling. However, the information of source and target domains is mixed in the domain-agnostic prompt. Hence, some results of DAPrompt-DS cannot match the performance of zero-shot CLIP.

5) Class-specific prompt may be useful. Through the above experiments, compared with unified context, class-specific prompts could improve the performance at around 0.1%–0.4% in some tasks.

6) Domain-specific prompt is critical to DAPrompt. Domain-specific prompt helps to learn the underlying label distribution of the target domain, which depicts the knowledge of target domain. Furthermore, domain-specific prompt allows for the dynamic mechanism to adapt the classifier according to the domain. We observe that domain-specific prompt improves the performance by 2.9%/1.3% and 1.3%/1.1% compared with DAPrompt-DS. Our DAPrompt achieves an accuracy of 86.9%/89.8%, which is 2.5%/1.1% higher than zero-shot CLIP (ResNet50/ViT-B/16).

2) Ablation on Pseudo Label Threshold: In Table I(b), we present the sensitivity of our method to the hyperparameter τ by ranging it from 0.4 to 0.7. It seems that our method is not sensitive to τ because of the tradeoff between quality and quantity of pseudo labels. For example, when τ is set to 0.7, the model is trained with fewer, but more confident pseudo labels and the quality of pseudo labels may make up for the performance drop brought by the reduced quantity.
3) Ablation on Context Token Length: We conduct experiments in Table I(c) to explore the influence of context token length. The lengths of domain-agnostic and domain-specific context tokens are denoted by $M_1$ and $M_2$, respectively. From the results, we can see that the performance is a little lower when $M_1 < M_2$. Overall, the token length has little effect on the performance of our method. This implies that the continuous representation could be learned with a small number of tokens.

4) Ablation on miniDomainNet and Office-Home: Considering that zero-shot CLIP is the best baseline model of these variants, we also compare zero-shot CLIP with our method DAPrompt on the miniDomainNet and Office-Home datasets in Table II. We also show the results of fine-tuning. ResNet50 and ViT-B/16 are selected as its visual encoder. The results show that on both datasets, fine-tuning of CLIP performs poorly, which further validates our previous observation. On the Office-Home dataset, our method outperforms CLIP by 2.5% / 2.0%. On a larger and more complicated dataset miniDomainNet, our method achieves an improvement of 3.6% / 3.0% over CLIP. Among these subtasks, our DAPrompt significantly improved the accuracy by 5.0% and 4.7% on S→P and S→C, respectively. Our method could perform well on both datasets with small size of examples (Office-Home) and large datasets with more than 100 categories (miniDomainNet). What is more, on all the 12 tasks in each dataset, DAPrompt consistently improves the performances. Overall, the results show that our DAPrompt is generic to different datasets.

D. Comparison With State-of-the-Art Methods

Considering that a few UDA methods are evaluated on miniDomainNet with ResNet50, we compare our DAPrompt with sota methods on Office-Home and VisDA2017 with ResNet50 and ResNet101 as image encoder, respectively, for a fair comparison.

1) Results on Office-Home: They are shown in Table III, where our method obviously outperforms all other baselines with respect to the average accuracy of 12 tasks. Note that there exists a large performance gap between the feature alignment-based methods (e.g., DANN [9] and CDAN + E [17]) and SRDC [21]. The possible reason may be that excessive feature alignment would hamper the discrimination of target data. While such potential risk will not happen in our...
method, we do not force feature alignment across domains. In particular, our method further surpasses the state-of-the-art method SRDC [21] by a large margin of 3.2% in terms of average accuracy. Compared with the baseline method CLIP, our method has improved all the 12 UDA tasks and outperforms it by 2.5% on the Office-Home dataset. We owe the performance improvement to the more suitable visual concepts for the target domain that are generated from our learned prompts. The superior performance of our method shows that simple prompt learning is effective for UDA problems.

### 2) Results on VisDA-2017 [29]:

They are presented in Table IV. It can be observed that our method achieves the highest average accuracy of 86.9% over the 12 classes, outperforming the state-of-the-art method STAR [19] by a large margin of 4.2%. Note that CLIP in Table IV means zero-shot CLIP, which adopts “a photo of a [CLASS]” as the handcrafted prompt. Even when the handcrafted prompt method already has an impressive performance, our DAPrompt still achieves a 2.5% absolute improvement over it. The reason why the accuracy of truck is significantly boosted may be that the concept of “truck” is more discriminative in the language model. Furthermore, with the help of prompt learning, DAPrompt outperforms CLIP by 7.5%, 14%, and 9.2% on “knife,” “person,” and “plant.” In general, despite the simplicity of our method, the encouraging results validate the efficacy of our prompt learning method.

### 3) Training Time Analysis:

To compare the training efficiency, we train all the models with one NVIDIA RTX 2080 Ti GPU. Our method is much more efficient than other methods. For example, MCD [38] and DANN [9] take 13.4 and 38.3 h to train on VisDA-2017, respectively, while our DAPrompt only costs 5.3 h. Because we only fine-tune the prompt with very few parameters, it is much easier and faster to optimize the model.

### E. Visualization

In Fig. 5, we compare the prediction confidence of the ground-truth category on the target domain when using three different prompting methods: 1) a handcrafted prompt (this method is identical to CLIP); 2) the prompt with only domain-agnostic context; and 3) the prompt with domain-agnostic context and domain-specific context (our DAPrompt).

For the first example, the plant only takes up a small area of the image. Hence, the prompt “a photo of a plant” is inappropriate for the image, while “a photo of a plant with a pot” might be a better match. Therefore, the handcrafted prompt performs poorly on this example. In contrast, the learnable prompt yields a more confident prediction than the manually designed prompt. For the last example, it is a good match for the prompt “a photo of a backpack.” The learnable domain-agnostic context even performs worse than the manually designed prompt. However, images from the “product” domain always share some common features.
Our experiments demonstrate that the major improvement the performance of strong baseline method by 2.5%, 2.5%, and have demonstrated the advantage of our method and it could prompt learning, we build a bridge between multimodality the classifier according to the sample. By making use of conditional probability of target domain and dynamically adapt to advocate learning distinct domain representations of the source and the target domain. Our method could learn the conditional probability of target domain and dynamically adapt the classifier according to the sample. By making use of prompt learning, we build a bridge between multimodality methods and domain adaptation methods. Extensive results have demonstrated the advantage of our method and it could serve as a good baseline for future work. DAPrompt improved the performance of strong baseline method by 2.5%, 2.5%, and 3.6% on Office-Home, VisDA-2017, and mini-DomainNet. Our experiments demonstrate that the major improvement comes from our core design domain-specific prompt.

V. CONCLUSION

In this article, we introduce a novel prompt learning method for unsupervised domain adaptation, which is free of aligning features and benefits from preserving semantic features. It is simple yet effective, easy to implement, and fast for training. We design domain-specific context for each domain to advocate learning distinct domain representations of the source and the target domain. Our method could learn the conditional probability of target domain and dynamically adapt the classifier according to the sample. By making use of prompt learning, we build a bridge between multimodality methods and domain adaptation methods. Extensive results have demonstrated the advantage of our method and it could serve as a good baseline for future work. DAPrompt improved the performance of strong baseline method by 2.5%, 2.5%, and 3.6% on Office-Home, VisDA-2017, and mini-DomainNet. Our experiments demonstrate that the major improvement comes from our core design domain-specific prompt.

REFERENCES

[1] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2016, pp. 770–778.
[2] G. Huang, Z. Liu, G. Pleiss, L. V. D. Maaten, and K. Q. Weinberger, “Convolutional networks with dense connectivity,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 44, no. 12, pp. 8704–8716, Dec. 2022.
[3] A. Radford et al., “Learning transferable visual models from natural language supervision,” in Proc. Int. Conf. Mach. Learn., vol. 139, 2021, pp. 8748–8763.
[4] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, “ImageNet: A large-scale hierarchical image database,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., Jun. 2009, pp. 248–255.
[5] S. Ben-David, J. Blitzer, K. Crammer, A. Kulesza, F. Pereira, and J. W. Vaughan, “A theory of learning from different domains,” Mach. Learn., vol. 79, nos. 1–2, pp. 151–175, May 2010.
[6] S. Ben-David, J. Blitzer, K. Crammer, and F. Pereira, “Analysis of representations for domain adaptation,” in Proc. NIPS, 2006, pp. 137–144.
[7] A. Torralba and A. A. Efros, “Unbiased look at dataset bias,” in Proc. CVPR, Jun. 2011, pp. 1521–1528.
[8] S. J. Pan and Q. Yang, “A survey on transfer learning,” IEEE Trans. Knowl. Data Eng., vol. 22, no. 10, pp. 1345–1359, Jan. 2010.
[9] Y. Ganin and V. Lempitsky, “Unsupervised domain adaptation by backpropagation,” in Proc. ICML, 2015, pp. 1180–1189.
[10] S. Li et al., “Discriminative transfer feature and label consistency for cross-domain image classification,” IEEE Trans. Neural Netw. Learn. Syst., vol. 31, no. 11, pp. 4842–4856, Nov. 2020.
[11] H. Rangwani, S. K. Aithal, M. Mishra, A. Jain, and R. V. Babu, “A closer look at smoothness in domain adversarial training,” in Proc. Int. Conf. Mach. Learn., 2022, pp. 18378–18399.
[12] E. Tzeng, J. Hoffman, N. Zhang, K. Saenko, and T. Darrell, “Deep domain confusion: Maximizing for domain invariance,” 2014, arXiv:1412.5474.
[13] M. Long, Y. Cao, J. Wang, and M. Jordan, “Learning transferable features with deep adaptation networks,” in Proc. ICML, 2015, pp. 97–105.
[14] W. Zellinger, T. Grubinger, E. Luhgofer, T. Natschlager, and S. Saminger-Platz, “Central moment discrepancy (CMD) for domain-invariant representation learning,” in Proc. ICLR, 2017.

Authorized licensed use limited to the terms of the applicable license agreement with IEEE. Restrictions apply.
[40] S. Li et al., “Domain conditioned adaptation network,” in Proc. AAAI Conf. Artif. Intell., 2020, vol. 34, no. 7, pp. 11386–11393.

[41] M. Li, Y.-M. Zhai, Y.-W. Luo, P.-F. Ge, and C.-X. Ren, “Enhanced transport distance for unsupervised domain adaptation,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2020, pp. 13933–13941.

[42] S. Cui, S. Wang, J. Zhao, L. Li, Q. Huang, and Q. Tian, “Towards discriminability and diversity: Batch-norm-norm maximization under label insufficient situations,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2020, pp. 3940–3949.

[43] C. Jia et al., “Scaling up visual and vision-language representation learning with noisy text supervision,” in Proc. ICML, 2019, pp. 4904–4916.

[44] F. Petroni et al., “Language models as knowledge bases?” in Proc. Conf. Empirical Methods Natural Lang. Process., 9th Int. Joint Conf. Natural Lang. Process. (EMNLP-IJCNLP), 2019, pp. 2463–2473.

[45] P. Liu, W. Yuan, J. Fu, Z. Jiang, H. Hayashi, and G. Neubig, “Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing,” 2021, arXiv:2107.13586.

[46] T. Shin, Y. Raegehi, R. L. Logan IV, E. Wallace, and S. Singh, “AutoPrompt: Eliciting knowledge from language models with automatically generated prompts,” in Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP), 2020, pp. 4222–4235.

[47] B. Lester, R. Al-Rfou, and N. Constant, “The power of scale for parameter-efficient prompt tuning,” 2021, arXiv:2104.08691.

[48] X. L. Li and P. Liang, “Prefix-tuning: Optimizing continuous prompts for generation,” in Proc. 59th Ann. Meeting Assoc. Comput. Linguistics, 11th Int. Joint Conf. Natural Lang. Process., 2021, pp. 4582–4597.

[49] Z. Jiang, F. F. Xu, J. Araki, and G. Neubig, “How can we know what language models know?”, Trans. Assoc. Comput. Linguistics, vol. 8, pp. 423–438, Jul. 2020.

[50] N. Feiner, U. Waltinger, and H. Schütte, “E-BERT: Efficient yet effective entity embeddings for BERT,” 2019, arXiv:1911.03681.

[51] K. Zhou, J. Yang, C. Change Loy, and Z. Liu, “Learning to prompt for vision-language models,” 2021, arXiv:2109.01134.

[52] A. Dosovitskiy et al., “An image is worth 16x16 words: Transformers for image recognition at scale,” in Proc. ICLR, 2021.

[53] A. Vaswani et al., “Attention is all you need,” in Proc. NIPS, 2017, pp. 5998–6008.

[54] T. Lin et al., “Microsoft COCO: Common objects in context,” in Proc. ECCV, vol. 8693, 2014, pp. 740–755.

[55] P. Xiong, Q. Bai, X. Xia, Z. Huang, K. Saenko, and B. Wang, “Moment matching for multi-source domain adaptation,” in Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV), Oct. 2019, pp. 1406–1415.

[56] K. Zhou, Y. Yang, Y. Qiao, and T. Xiang, “Domain adaptive ensemble learning,” IEEE Trans. Image Process., vol. 30, pp. 8008–8018, 2021.

[57] I. Loshchilov and F. Hutter, “SGDR: Stochastic gradient descent with warm restarts,” in Proc. ICLR, 2017.

[58] Y. Zhang, H. Tang, K. Jia, and M. Tan, “Domain-symmetric networks for adversarial domain adaptation,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2019, pp. 5026–5035.

[59] L. Hu, M. Kan, S. Shan, and X. Chen, “Unsupervised domain adaptation with hierarchical gradient synchronization,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2020, pp. 4042–4051.

[60] S. Cui, S. Wang, J. Zhao, C. Su, Q. Huang, and Q. Tian, “Gradually vanishing bridge for adversarial domain adaptation,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2020, pp. 12452–12461.

[61] X. Gu, J. Sun, and Z. Xu, “Spherical space domain adaptation with robust pseudo-label loss,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2019, pp. 9098–9107.

[62] Y. Zhang et al., “Free lunch for domain adversarial training: Environment label smoothing,” 2023, arXiv:2302.00194.

[63] C.-Y. Lee, T. Batra, M. H. Baig, and D. Ullbricht, “Sliced Wasserstein discrepancy for unsupervised domain adaptation,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2019, pp. 10272–10281.

[64] R. Li, Q. Jiao, W. Cao, H.-S. Wong, and S. Wu, “Model adaptation: Unsupervised domain adaptation without source data,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2020, pp. 9638–9647.

[65] Z. Du, J. Li, H. Su, L. Zhu, and K. Lu, “Cross-domain gradient discrepancy minimization for unsupervised domain adaptation,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2021, pp. 3937–3946.