COSMo: CLIP Talks on Open-Set Multi-Target Domain Adaptation

Munish Monga[1]
munish30monga@gmail.com
Sachin Kumar Giroh[1]
22m2159@iitb.ac.in
Ankit Jha[1,2]
ankitjha16@gmail.com
Mainak Singha[1]
mainaksingha.iitb@gmail.com
Biplab Banerjee[1]
getbiplab@gmail.com
Jocelyn Chanussot[2]
jocelyn.chanussot@inria.fr

[1] Indian Institute of Technology, Bombay
Mumbai, India
[2] INRIA, Grenoble, France

Abstract

Multi-Target Domain Adaptation (MTDA) entails learning domain-invariant information from a single source domain and applying it to multiple unlabeled target domains. Yet, existing MTDA methods predominantly focus on addressing domain shifts within visual features, often overlooking semantic features and struggling to handle unknown classes, resulting in what is known as Open-Set (OS) MTDA. While large-scale vision-language foundation models like CLIP show promise, their potential for MTDA remains largely unexplored. This paper introduces COSMo, a novel method that learns domain-agnostic prompts through source domain-guided prompt learning to tackle the MTDA problem in the prompt space. By leveraging a domain-specific bias network and separate prompts for known and unknown classes, COSMo effectively adapts across domain and class shifts. To the best of our knowledge, COSMo is the first method to address Open-Set Multi-Target DA (OSMTDA), offering a more realistic representation of real-world scenarios and addressing the challenges of both open-set and multi-target DA. COSMo demonstrates an average improvement of 5.1% across three challenging datasets: Mini-DomainNet, Office-31, and Office-Home, compared to other related DA methods adapted to operate within the OSMTDA setting.

1 Introduction

Domain adaptation (DA) techniques play a crucial role in improving the generalizability of machine learning models across diverse data distributions. DA aims to address the domain shift problem, where models trained on one domain may struggle to generalize effectively to another related domain. Traditionally, DA has mainly focused on a closed-set (CS) setting [8, 14, 16, 33], assuming that the classes in the source and target domains are identical. However, this assumption often does not hold in practical applications, especially in open-world scenarios where class shifts, i.e., unknown classes in the target domain, may occur.

© 2024. The copyright of this document resides with its authors.
It may be distributed unchanged freely in print or electronic forms.

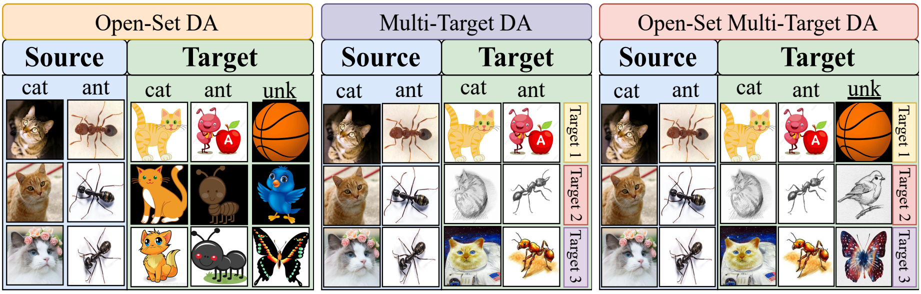


Figure 1: OSMTDA differs from traditional DA settings like Open-set DA by handling unknown classes across diverse target domains, while Multi-target DA transfers knowledge from a single labeled source to multiple unlabeled targets. unk denotes the unknown class.

Open-set DA (OSDA) [3, 5, 23, 28, 35] extends the CS framework by considering classes present in the target domain but absent in the source domain, making it more realistic .

Single-Target DA (STDA) [26, 27, 44] deals with adapting a model from a single source domain to a single target domain. While effective in controlled settings, STDA encounters challenges in scalability and handling diversity in real-world applications, especially when facing multiple domain shifts simultaneously. Transitioning to a more intricate scenario, Multi-Target DA (MTDA) focuses on adapting a model to perform well across multiple target domains. In situations where domain labels are absent, Blended-Target DA (BTDA) methods [43] merge all targets into one for source-target alignment.

In this paper, we introduce Open-set Multi-Target Domain Adaptation (OSMTDA, as shown in Figure 1), merging elements of OSDA and MTDA. OSMTDA tackles three key challenges: managing domain shifts across multiple targets, handling class shifts in open-world scenarios, and adapting from one source to multiple targets. This framework is novel and unexplored. OSMTDA is crucial in domains like autonomous driving, ensuring adaptability across diverse contexts. For instance, vehicles trained in one area may encounter new traffic conditions elsewhere. OSMTDA addresses real-world challenges without extensive labeling or fine-tuning. Additionally, it holds promise in federated learning, bolstering model adaptability while preserving data privacy.

Recently, Vision Language Models (VLMs) [20, 31], which are trained on a huge amount of data, have shown impressive performance over a wide range of tasks, but they still underperform in downstream tasks. Various techniques are used to make VLMs perform better, such as feature adapter [11], Prompt Learning [45, 46] etc. Textual Prompt Learning in VLMs is effectively utilised in CoOp [46] where the prompts that are being fed to the text encoder are made learnable, CoCoOp [45] adds on conditional bias, [25] utilised visual prompts. Maple [21] utilises both the text and image prompts. Domain Adaptation via Prompt Learning [12] utilised prompt learning to solve the domain adaptation task; however, it utilises the domain labels.

In this paper, we address the OSMTDA problem using source domain-guided prompt learning in VLMs. By combining prompt learning with a domain-specific bias network, we extract knowledge from the source domain.We incorporate domain-specific bias into separate learnable prompts for known and unknown classes to improve the alignment of image and text pairs (discussed in detail in Section 3). In OSMTDA, we have labeled data from the source domain and unlabeled data from target domains, which include the known classes from the source domain and unlabeled classes as unknown classes. We adopt the similar approach to BTDA [7], merging all targets into one and assuming that domain labels are

unavailable. We highlight our contributions as:
-To the best of our knowledge, our method is the first to address the task of open-set domain adaptation for multi-target scenarios.
-We propose a source domain-guided prompt learning approach. Separate prompts for known and unknown classes handle the class shift, while a domain-specific bias network addresses the domain shift.
-COSMo outperforms the referred DA methods adapted for OSMTDA, demonstrating significant performance improvement atleast by 5.1% across the Office-Home, Office31, and Mini DomainNet datasets.

2 Related Work

2.1 Open-set Domain Adaptation

In domain adaptation, the concept of open-set domain adaptation (OSDA) was introduced by [5]. OSDA employs alignment techniques to align the feature spaces of the source and target domains. However, this method relies on unknown source labels, potentially distorting semantic features crucial for class differentiation [6]. In contrast, OSDA-BP [35] proposes an alternative approach that dispenses with unknown source labels and utilizes adversarial training. This technique enhances the model's ability to distinguish between known and unknown target samples by learning discriminative features invariant across domains. Traditional domain adaptation strategies, such as distribution matching and extracting domain-invariant features, often utilize metrics like Maximum Mean Discrepancy (MMD) to measure domain distances. However, these methods typically overlook the possibility of encountering examples from unknown classes in the target domain, limiting their applicability in open-set scenarios. Additionally, DANCE [36] presents another notable approach, leveraging neighborhood clustering and entropy-based feature alignment to address the challenges of universal domain adaptation.

2.2 Multi-target domain adaptation

Multi-Target Domain Adaptation (MTDA) aims to bridge the domain gap by transferring knowledge from a single source domain. This setting has been extensively explored across various tasks, including classification, segmentation [18, 37], and object detection. In our research, we adopt a Blended Multi-Target Domain Adaptation approach akin to the framework presented in [43], as we considered the unavailability of domain labels.

Significant efforts in MTDA include using information-theoretic strategies to segregate shared and private information across domains, as implemented in [13]. Additionally, [17] addresses multi-target domain adaptation for segmentation tasks through a collaborative learning framework. Common strategies for MTDA include adversarial learning, which leverages adversarial networks to minimize domain discrepancies; graph-based methods [32], which utilize Graph Convolution Networks (GCN) to exploit relational data within and across domains; and knowledge distillation techniques. *In our proposed COSMo, We utilize source domain guided prompt learning to segregate the sets of known and unknown classes in OSMTDA.*

2.3 Vision-language models and prompt learning

Multi-modal vision-language models have made significant strides in various image recognition tasks, utilizing advanced language models like BERT [9] and GPT [30] alongside CNN and ViT for visual analysis. Notable examples include CLIP [31] and VisualBERT

[24]. Traditionally, these models relied on manually crafted textual prompts, which can be complex. Prompt learning methods [4, 21, 38, 45, 46] have gained traction for effectively tailoring prompts for downstream tasks. These methods treat token embeddings as learnable variables constrained by image features. Recently, CLIP has been utilized to address challenges in domain adaptation (DA) [39] and domain generalization (DG) [1, 40] tasks. DAPL create disentangled category (class) and domain representations by aligning them differently. [12] employs domain-specific context tokens for unsupervised DA, while AD-CLIP [39] generates domain-agnostic tokens via a cross-domain style mapping network inspired by STYLIP [1]. CLIPN [42] tackles out-of-distribution (OOD) tasks by training a "no" text encoder for negative semantic prompts in addition to positive ones. However, these methods are explicitly designed for DG, DA, and OOD tasks. Our model differs from DAPL in the prompt learning technique, also in DAPL, the domain labels are assumed to be known. *In this paper, we utilize CLIP with prompt learning for OSMTDA.*

3 Methodology

3.1 Problem formulation

In the context of OSMTDA, we possess labeled data, denoted as X_s, from a single source domain S, represented as $(X_s, Y_s) \in S$, and unlabeled data, X_{t_i}, from multiple target domains T_i, where $X_t = \bigcup_{i=1}^{q} X_{t_i}$ with $X_{t_i} \in T_i$, and q denotes the number of target domains. Here, C_s and C_t refer to the source and target domain classes, respectively. To establish an open-set scenario, we assume $C_s \subset C_t$, designating the known classes as those from the source domain, denoted as $C_k = C_s$, while the target domain may contain additional unknown classes, denoted as $C_u = C_t \setminus C_s$. For a multi-target setup, we ensure that the target domain classes C_t remain consistent across all q target domains, i.e., $C_t = \{C_{t_i}\}_{i=1}^{q}$. During training, a mini-batch of size N comprises two sets of data: $D_s = (x_s^i, y_s^i)_{i=1}^{N}$ sampled from (X_s, Y_s) and $D_t = (x_t^i)_{i=1}^{N}$ sampled from X_t. For a given target image x_t from any target domain, the objective in OSMTDA is to accurately classify x_t into one of the known classes C_k or identify it as an `unknown` class.

3.2 Overview of COSMo

In COSMo, we aim to learn domain-specific and domain-agnostic information. Domain-specific information is learnt via the Domain-Specific Bias Network (DSBN), $(B_\theta(\cdot))$ and the domain-agnostic information is learnt via the separate learnable prompts (known and unknown class-based prompts) for the known and unknown classes. DSBN is trained on both the source and target domain instances, whereas the known and unknown class-based prompts are trained on D_s and D_t, respectively, as shown in Figure 2. We leverage the pre-trained vision encoder $\mathcal{F}_v$ and text encoder $\mathcal{F}_t$ of CLIP. We provide detailed description on the components and working of our proposed COSMo in the following paragraphs.

Domain-Specific Bias Network (DSBN): DSBN captures domain-specific information from the image features and helps address the domain distribution shift. DSBN is parameterized by θ, and it modifies the learnable prompts, as the output of B_θ is directly added to the learnable prompts. The domain information helps in better alignment of the image and text embeddings, as text embedding is based on the unique characteristics of each domain, thus improving the model's adaptability across various domains.

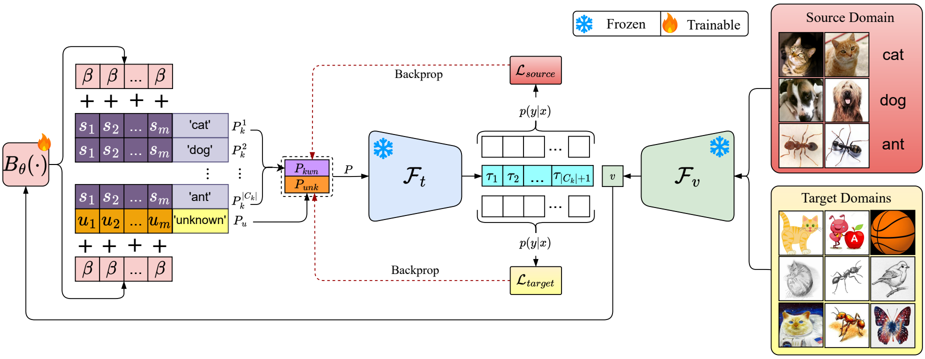


Figure 2: The architecture overview of COSMo, where $\mathcal{F}_v$ and $\mathcal{F}_t$ are the frozen pretrained CLIP's image and text encoders, respectively. P_{kwn} and P_{unk} denote the prompts for the known and unknown classes, respectively. $\mathcal{B}_\theta(\cdot)$ represents the domain specific bias network, which generates the domain-bias context tokens β. Best view in color.

Source Domain-Guided Prompt Learning (SDGPL): Our approach employs a source domain-guided prompt learning strategy in the prompt space, utilizing a different prompt for the known and unknown classes. The prompts, thus trained, are domain agnostic, and the domain bias is added via the domain-specific bias network.

- **Known class prompts** (P_{kwn}): Trained on the instances of the source domain classes (C_s). As depicted in figure 2, P_k^c captures the domain-agnostic information for class c, where $c \in C_s$ and helps align image-text embedding pairs of the known classes. The known class prompts are constructed as follows:

$$P_k^c = [s_1][s_2]\ldots[s_m][\text{CLS}]_c, \quad P_{k,\ bias}^c = [s_1+\beta][s_2+\beta]\ldots[s_m+\beta][\text{CLS}]_c, \tag{1}$$

$$P_{kwn} = [P_{k,\ bias}^1; P_{k,\ bias}^2; \ldots ; P_{k,\ bias}^{|C_k|}] \tag{2}$$

  where s_i, for $i \in \{1,\ldots,m\}$, represents the i^{th} context vector (learnable component) of the known class-based prompts and is the same for all the classes in C_k, m denotes the length of the context prompt, β denotes the domain-bias context token obtained from the domain-specific bias network (B_θ), $P_{k,\ bias}^c$ represents the biased known class prompt, P_{kwn} represents the cumulative prompt for $|C_k|$ classes, and $[\text{CLS}]_c$ denotes the class vector for class $c \in C_k$.

- **Unknown class prompts** (P_{unk}): Employed for adapting to and categorizing unknown classes in the target domains. As depicted in figure 2, unlike P_k, P_u is updated through the target domain instances by utilizing the pseudo labels obtained through the gained knowledge on the source domain, thereby enhancing the model's capability to effectively recognize new, unseen categories. The unknown class prompts are constructed as follows:

$$P_u = [u_1][u_2]\ldots[u_m][\text{UNK}], \quad P_{unk} = [u_1+\beta][u_2+\beta]\ldots[u_m+\beta][\text{UNK}] \tag{3}$$

  where u_i represents the context components of the unknown class prompts P_{unk}, m is the length of the prompt, and [UNK] denotes the class vector corresponding to the token "*unknown.*".

Algorithm 1 Pseudo code to train COSMo

```
function MODEL($x$)
    $v \leftarrow \mathcal{F}_v(x)$
    $\beta \leftarrow B_\theta(v)$
    Get $P_{kwn}$ using $s_i$ and $\beta$, and $P_{unk}$ using $u_i$ and $\beta$    ▷ See equations 1 and 3
    $\tau \leftarrow \mathcal{F}_t([P_{kwn}; P_{unk}])$
    $logits \leftarrow \tau_s^T \cdot v$
    return $logits$
end function

for each $(x_s, y_s)$ in $D_s$ and $x_t$ in $D_t$ do
    $\text{logits}_s \leftarrow$ MODEL$(x_s)$    ▷ Freeze $u_i \; \forall \; i \in \{1, \dots, m\}$
    $\mathcal{L}_{\text{source}} \leftarrow \text{Loss}(\text{logits}_s, y_s)$    ▷ Update $\theta$ and $s_i \; \forall \; i \in \{1, \dots, m\}$
    Unfreeze $u_i \; \forall \; i \in \{1, \dots, m\}$    ▷ Unfreeze $u_i$ for target training
    $y_t \leftarrow \text{get\_pseudo\_labels}(x_t)$    ▷ Get pseudo labels for target data
    $\text{logits}_t \leftarrow$ MODEL$(x_t)$    ▷ Freeze $s_i \; \forall \; i \in \{1, \dots, m\}$
    $\mathcal{L}_{\text{target}} \leftarrow \text{Loss}(\text{logits}_t, y_t)$    ▷ Update $\theta$ and $u_i \; \forall \; i \in \{1, \dots, m\}$
    Unfreeze $s_i \; \forall \; i \in \{1, \dots, m\}$    ▷ Unfreeze $s_i$ for next iteration
end for
```

3.3 Model optimization

In our CLIP-based model, an image x (from either the source or target domain) is processed by the image encoder $\mathcal{F}_v$, resulting in a feature vector v = $\mathcal{F}_v(x)$ and bias ($\beta = B_\theta(\mathcal{F}_v(x))$) is obtained via DSBN. Text prompts with bias (P_{knw} *and* P_{unk}) are encoded by the text encoder $\mathcal{F}_{txt}$ to get text features W containing the vectors $\tau_1, \tau_2, \dots, \tau_{|C_k|+1}$, where each vector corresponds to the encoded textual representation of a class (including the additional unknown class) with domain bias. The probability of an input image x belonging to class c, is computed as:

$$p(y_c|x) = \frac{\exp(\text{sim}(\tau_c, v)/\eta)}{\sum_{i=1}^{|C_k|+1} \exp(\text{sim}(\tau_i, v)/\eta)} \tag{4}$$

where $\text{sim}(\tau_c, v)$ represents the similarity between the image feature v and the text feature for class c is denoted by τ_c, and η is a temperature parameter that scales the logits before applying the softmax function. The cross-entropy loss for source domain instances, incorporating minimum entropy regularization, is calculated as follows:

$$\mathcal{L}_{\text{source}} = -\mathbb{E}_{(x,y)\sim p(X_S,Y_S)}[\log p(y|x)] + \lambda \cdot \mathbb{E}_{x\sim p(X_S)}\left[-\sum_{c=1}^{|C_k|+1} p(y_c|x)\log p(y_c|x)\right] \tag{5}$$

Similarly, we define $\mathcal{L}_{\text{target}}$ as the target loss, which consists of cross entropy and entropy regularization losses, defined in Eq. 6. Here, $\tilde{y}_c$ represents the obtained pseudo label, λ is a hyperparameter controlling the strength of the entropy regularization. Minimum Entropy regularisation ensures that the model gives prediction with high confidence, whereas the cross entropy loss ensures that the model gives correct prediction.

$$\mathcal{L}_{\text{target}} = -\mathbb{E}_{(x,\tilde{y})\sim p(X_T,\tilde{Y}_T)}[\log p(y|x)] + \lambda \cdot \mathbb{E}_{x\sim p(X_T)}\left[-\sum_{c=1}^{|C_k|+1} p(\tilde{y}_c|x)\log p(\tilde{y}_c|x)\right] \tag{6}$$

The pseudo-labels are assigned based on the confidence thresholds. More specifically, for an unlabeled instance, if the $p(y_c|x)$ (probability of an input image x belonging to class c) for each known class is less than κ_{lower}, then the instance is labelled as an unknown. Also, if the maximum probability for an unknown class is at least κ_{upper}, then the instance is labelled as unknown. This selective approach allows the model to focus on more reliable, high-confidence predictions and enhances the overall robustness and accuracy of the domain adaptation process. The training procedure is mentioned in Algorithm 1. During inference, classification is done with the learnt known and unknown class prompt, assigning the class corresponding to the maximum probability value $(p(y_c|x))$.

4 Experiments

Datasets: For OSMTDA, we selected three widely used datasets: Office-31 [34], Office-Home [41], and Mini-DomainNet [29]. **Office-31** contains 31 classes across three domains; Amazon (A), DSLR (D), Webcam (W), while **Office-Home** has 65 classes spanning Art (A), Clip Art (C), Product (P), and Real World (R) domains. **Mini-DomainNet**, a subset of DomainNet, features 126 classes distributed among Clipart (C), Painting (P), Real (R), and Sketch (S) domains. We ensure an open-set setting by dividing classes into known and unknown categories, maintaining a ratio of $|C_k|/|C_u| = 10/21$ for Office-31, $15/50$ for Office-Home, and $60/66$ for Mini-DomainNet.

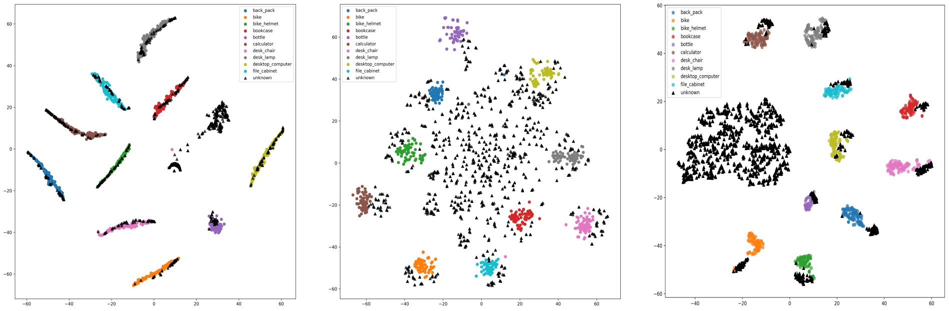
Experimental Details and Evaluation Metrics: We use the *AdamW* optimizer [22] to optimize COSMo, with a batch size of 32 and used cosine annealing with an initial learning rate of 0.001. We utilize two pre-trained vision encoders as $\mathcal{F}_v$: ViT-B/16 [10] (Table 1) and ResNet-50 [15] (results in `supplementary`). When assessing open-set multi-target domain adaptation, we used commonly employed metrics [2, 19]: average known class accuracy (OS*), the accuracy of unknown classes (UNK), and the harmonic mean score (HOS) between the accuracy of known (OS*) and unknown classes (UNK). In `supplementary`, we provide the detail explanation of hyperparameter settings of our proposed COSMo.

4.1 Comparison to the literature

To the best of our knowledge, our proposed COSMo is the first attempt at addressing the challenges of OSMTDA, leaving no existing baselines for comparison. Hence for the proposed OSMTDA task, as shown in Table 1, we comprehensively compare COSMo's performance across three distinct datasets against four relevant baseline models: (1) the zero-shot predictions by CLIP [31] were taken by using standard prompts like: "a dog", "a cat" where class names were provided for the source classes and keeping a $|C_k|$ class classifier, to ensure open-set setting if the maximum probability of predicted class is less than a threshold, it was predicted as unknown. (2) OSDA-BP [5] and (3) DANCE [36] are used as non-CLIP baselines. Since these methods work under the OSDA setting but with a single target, we merged all target domains into one. (4) AD-CLIP [39] also tackles the DA problem with CLIP; but works in a closed-set setup; to ensure open-set setting; we employed the threshold technique as CLIP. COSMo consistently outperforms these methods in terms of the HOS metric, surpassing CLIP, OSDA-BP, DANCE, and AD-CLIP, demonstrating an average improvement of 52.7%, 54.6%, 5.1%, and 83.3%, respectively, averaged across all three datasets. We provide detailed results for various domain permutations within each dataset in the `supplementary` material.

Table 1: Comparison of OSMTDA task with our proposed COSMo with state-of-the-art on the Office-31, Office-Home and Mini-DomainNet datasets. The best results are highlighted in **bold**. We report the HOS metric score. Results are reported using ViT-B/16 backbone.

	Office-31				**Office-Home**					**Mini-DomainNet**				
Methods	**A**	**D**	**W**	**Avg.**	**A**	**C**	**P**	**R**	**Avg.**	**C**	**P**	**R**	**S**	**Avg.**
CLIP [31]	43.79	39.95	39.61	41.12	61.76	61.16	67.24	63.52	63.42	75.78	75.67	72.39	75.66	74.88
OSDA-BP [35]	83.92	74.60	73.17	77.23	54.39	33.01	55.97	56.95	50.08	41.27	52.41	48.86	37.79	45.08
DANCE [36]	86.48	86.34	87.76	86.86	**81.20**	85.42	78.58	79.28	81.12	70.26	81.29	73.35	73.84	74.69
AD-CLIP [39]	36.97	48.81	32.32	39.37	54.04	48.91	49.67	46.69	49.83	53.16	56.92	47.61	53.88	52.89
COSMo (Ours)	**92.46**	**88.41**	**89.15**	**90.01**	80.96	**86.8**	**82.42**	**81.97**	**83.04**	**81.05**	**84.15**	**79.29**	**82.89**	**81.84**



(a) OSDA-BP (b) DANCE (c) COSMo (Ours)

Figure 3: t-SNE visualizations on the Office31 Dataset with Amazon as the source domain. Colored dots represent known classes in the source domain, while black triangles denote target domain samples. For COSMo, text embeddings are used, while features from the penultimate layer are used for the other models.

Furthermore, we visualize the t-SNE embeddings generated from the text encoder of our proposed COSMo for both known and unknown classes, as depicted in Figure 3. It is evident that COSMo effectively distinguishes between known and unknown classes across diverse target domains, showcasing superior segregation compared to OSDA-BP [35] and DANCE [36]. This observation highlights the robustness of our proposed COSMo in effectively handling both familiar and novel data instances.

4.2 Ablation studies

Impact of having separate P_{kwn} and P_{unk}: Here, we discuss the necessity of having separate prompts for known and unknown classes in our proposed COSMo for the OSMTDA task. From Table 2, it's evident that COSMo with individual prompts for known and unknown classes improves by a margin of 0.38% on the HOS metric. However, experimental findings reveal that using common prompts for known and unknown classes leads to model overfitting to unknown classes, significantly decreasing overall performance. Hence, employing separate prompts plays a crucial role in enhancing the model's ability to effectively adapt to open-set scenarios, thereby improving its overall performance. Further details on the ablations are discussed in the `supplementary` material.

Role of Domain-Specific Bias Network (DSBN): Additionally, we conduct an ablation study on COSMo, integrating the DSBN network alongside separate prompting for known and unknown classes. Table 2 illustrates the impact, showing a significant enhancement in addressing domain shifts for the OSMTDA task. Specifically, the inclusion of DSBN results in an approximate 4% improvement in the HOS metric, underscoring its effectiveness in mitigating domain-related challenges.

Table 2: Ablations of our proposed COSMo with different components on the Office31 dataset with Amazon as source domain with ViT-B/16 backbone and context length m= 4.

Separate Prompt	DSBN	Entropy Regularisation	HOS
×	×	×	84.41
✓	×	×	84.79
✓	✓	×	88.59
✓	✓	✓	**92.46**

Table 3: Ablations of our proposed COSMo with different context lengths with ViT-B/16 backbone.

Context length (m)	Trainable Parameters (in K)	HOS
4	37.4	**92.46**
8	41.5	88.83
16	49.7	89.58

Ablation with Entropy Regularization loss: Finally, we ablate our proposed COSMo model with the loss terms discussed in Eq. 5 and 6. Initially, we train the network using only the cross-entropy loss, which fails to classify open-set samples from unseen target domains. Subsequently, we incorporate the entropy regularization loss alongside the cross-entropy loss and train COSMo. Table 2 demonstrates that it achieves approximately 4% better HOS performance compared to optimizing the network solely with cross-entropy loss.

Ablation with number of context tokens and number of trainable parameters: We examine the impact of prompt learning in COSMo by varying the context token length with $m = 4, 8$, and 16. Table 2 shows the HOS metric performance on the OSMTDA task, with COSMo trained on the Amazon domain and evaluated on other domains of the Office31 dataset. Notably, $m = 4$ achieves superior performance by a minimum margin of 2.88% across all settings detailed in Table 3. Additionally, we report the number of trainable parameters required for training COSMo, observing that with $m = 4$, COSMo achieves higher performance with fewer trainable parameters compared to $m = 8$ and 16.

5 Conclusion

In this paper, we introduce the novel framework COSMo, addressing the problem of Open-Set Multi-Target Domain Adaptation by leveraging source domain-guided prompt learning. We utilize the frozen image and text encoders of the pre-trained CLIP, along with a few trainable parameters, in designing the network. COSMo incorporates a domain-specific bias network and separate prompts for known and unknown classes, enabling efficient adaptation across domain and class shifts. To our knowledge, we are the first to tackle the challenges of the OSMTDA task, providing a practical representation of real-world scenarios by integrating both open-set and multi-target domain adaptation challenges. While COSMo assumes a consistent set of target domain classes across all domains, real-world scenarios may involve varying sets of classes across target domains. In future work, we aim to extend COSMo to address related tasks such as semantic segmentation and object detection.

6 Acknowledgements

We gratefully acknowledge the Centre for Machine Intelligence and Data Science (C-MInDS) at the Indian Institute of Technology Bombay for providing the financial support and infrastructure necessary to conduct this research. We also wish to express our appreciation for the valuable feedback provided by the reviewers, which significantly contributed to improving the quality of this paper.

References

[1] Shirsha Bose, Ankit Jha, Enrico Fini, Mainak Singha, Elisa Ricci, and Biplab Banerjee. Stylip: Multi-scale style-conditioned prompt learning for clip-based domain generalization. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 5542–5552, 2024.

[2] Silvia Bucci, Mohammad Reza Loghmani, and Tatiana Tommasi. On the effectiveness of image rotation for open set domain adaptation, 2020.

[3] Silvia Bucci, Mohammad Reza Loghmani, and Tatiana Tommasi. On the effectiveness of image rotation for open set domain adaptation. In *European conference on computer vision*, pages 422–438. Springer, 2020.

[4] Adrian Bulat and Georgios Tzimiropoulos. Lasp: Text-to-text optimization for language-aware soft prompting of vision & language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23232–23241, 2023.

[5] Pau Panareda Busto and Juergen Gall. Open set domain adaptation. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 754–763, 2017. doi: 10.1109/ICCV.2017.88.

[6] Ruichu Cai, Zijian Li, Pengfei Wei, Jie Qiao, Kun Zhang, and Zhifeng Hao. Learning disentangled semantic representation for domain adaptation. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence*, IJCAI-2019. International Joint Conferences on Artificial Intelligence Organization, August 2019. doi: 10.24963/ijcai.2019/285. URL http://dx.doi.org/10.24963/ijcai.2019/285.

[7] Ziliang Chen, Jingyu Zhuang, Xiaodan Liang, and Liang Lin. Blending-target domain adaptation by adversarial meta-adaptation networks, 2019.

[8] Sumit Chopra, Suhrid Balakrishnan, and Raghuraman Gopalan. Dlid: Deep learning for domain adaptation by interpolating between domains. In *ICML workshop on challenges in representation learning*, volume 2. Citeseer, 2013.

[9] 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.

[10] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[11] Peng Gao, Shijie Geng, Renrui Zhang, Teli Ma, Rongyao Fang, Yongfeng Zhang, Hongsheng Li, and Yu Qiao. Clip-adapter: Better vision-language models with feature adapters. *arXiv preprint arXiv:2110.04544*, 2021.

[12] 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.

[13] Behnam Gholami, Pritish Sahu, Ognjen Rudovic, Konstantinos Bousmalis, and Vladimir Pavlovic. Unsupervised multi-target domain adaptation: An information theoretic approach. *IEEE Transactions on Image Processing*, 29:3993–4002, 2020. doi: 10.1109/TIP.2019.2963389.

[14] Raghuraman Gopalan, Ruonan Li, and Rama Chellappa. Domain adaptation for object recognition: An unsupervised approach. In *2011 international conference on computer vision*, pages 999–1006. IEEE, 2011.

[15] 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.

[16] Tzu Ming Harry Hsu, Wei Yu Chen, Cheng-An Hou, Yao-Hung Hubert Tsai, Yi-Ren Yeh, and Yu-Chiang Frank Wang. Unsupervised domain adaptation with imbalanced cross-domain data. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4121–4129, 2015.

[17] Takashi Isobe, Xu Jia, Shuaijun Chen, Jianzhong He, Yongjie Shi, Jianzhuang Liu, Huchuan Lu, and Shengjin Wang. Multi-target domain adaptation with collaborative consistency learning, 2021.

[18] Takashi Isobe, Xu Jia, Shuaijun Chen, Jianzhong He, Yongjie Shi, Jianzhuang Liu, Huchuan Lu, and Shengjin Wang. Multi-target domain adaptation with collaborative consistency learning, 2021.

[19] JoonHo Jang, Byeonghu Na, DongHyeok Shin, Mingi Ji, Kyungwoo Song, and Il-Chul Moon. Unknown-aware domain adversarial learning for open-set domain adaptation, 2022.

[20] 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.

[21] Muhammad Uzair Khattak, Hanoona Rasheed, Muhammad Maaz, Salman Khan, and Fahad Shahbaz Khan. Maple: Multi-modal prompt learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 19113–19122, June 2023.

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

[23] Jogendra Nath Kundu, Naveen Venkat, Ambareesh Revanur, R Venkatesh Babu, et al. Towards inheritable models for open-set domain adaptation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12376–12385, 2020.

[24] 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.

[25] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning, 2023.

[26] Mingsheng Long, Han Zhu, Jianmin Wang, and Michael I Jordan. Unsupervised domain adaptation with residual transfer networks. *Advances in neural information processing systems*, 29, 2016.

[27] Mingsheng Long, Han Zhu, Jianmin Wang, and Michael I Jordan. Deep transfer learning with joint adaptation networks. In *International conference on machine learning*, pages 2208–2217. PMLR, 2017.

[28] Pau Panareda Busto and Juergen Gall. Open set domain adaptation. In *Proceedings of the IEEE international conference on computer vision*, pages 754–763, 2017.

[29] Xingchao Peng, Qinxun Bai, Xide Xia, Zijun Huang, Kate Saenko, and Bo Wang. Moment matching for multi-source domain adaptation, 2019.

[30] Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. Improving language understanding by generative pre-training. *https://www.mikecaptain.com/resources/pdf/GPT-1.pdf*, 2018.

[31] 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.

[32] Subhankar Roy, Evgeny Krivosheev, Zhun Zhong, Nicu Sebe, and Elisa Ricci. Curriculum graph co-teaching for multi-target domain adaptation, 2021.

[33] Kate Saenko, Brian Kulis, Mario Fritz, and Trevor Darrell. Adapting visual category models to new domains. In *Computer Vision–ECCV 2010: 11th European Conference on Computer Vision, Heraklion, Crete, Greece, September 5-11, 2010, Proceedings, Part IV 11*, pages 213–226. Springer, 2010.

[34] Kate Saenko, Brian Kulis, Mario Fritz, and Trevor Darrell. Adapting visual category models to new domains. In *European conference on computer vision*, pages 213–226. Springer, 2010.

[35] Kuniaki Saito, Shohei Yamamoto, Yoshitaka Ushiku, and Tatsuya Harada. Open set domain adaptation by backpropagation, 2018.

[36] Kuniaki Saito, Donghyun Kim, Stan Sclaroff, and Kate Saenko. Universal domain adaptation through self-supervision. 2020.

[37] Antoine Saporta, Tuan-Hung Vu, Matthieu Cord, and Patrick Pérez. Multi-target adversarial frameworks for domain adaptation in semantic segmentation, 2021.

[38] Mainak Singha, Ankit Jha, and Biplab Banerjee. Gopro: Generate and optimize prompts in clip using self-supervised learning. *arXiv preprint arXiv:2308.11605*, 2023.

[39] Mainak Singha, Harsh Pal, Ankit Jha, and Biplab Banerjee. Ad-clip: Adapting domains in prompt space using clip. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4355–4364, 2023.

[40] Mainak Singha, Ankit Jha, Shirsha Bose, Ashwin Nair, Moloud Abdar, and Biplab Banerjee. Unknown prompt the only lacuna: Unveiling clip's potential for open domain generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13309–13319, 2024.

[41] 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.

[42] Hualiang Wang, Yi Li, Huifeng Yao, and Xiaomeng Li. Clipn for zero-shot ood detection: Teaching clip to say no, 2023.

[43] Pengcheng Xu, Boyu Wang, and Charles Ling. Class overwhelms: Mutual conditional blended-target domain adaptation, 2023.

[44] 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.

[45] 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.

[46] 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.