AD-CLIP: Adapting Domains in Prompt Space Using CLIP

Mainak Singha Harsh Pal Ankit Jha Biplab Banerjee

Indian Institute of Technology Bombay

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.

Motivation

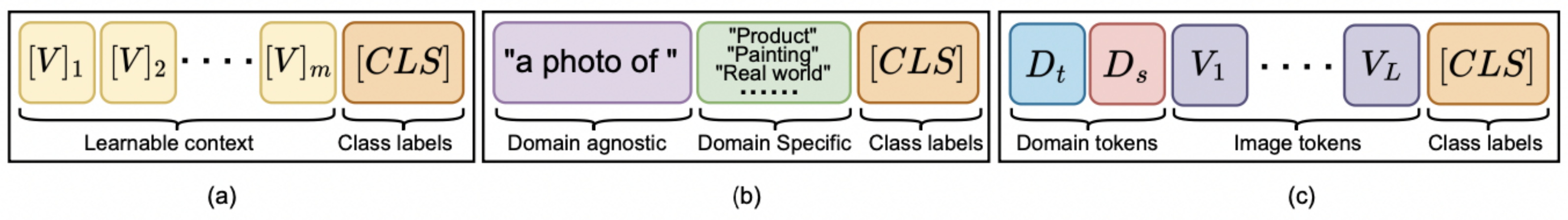


Figure 1. We highlight the differences between our prompts from the literature. a) CoOp [?] directly learns the prompt tokens from random vectors and may not be suitable for DA as it does not concern domain independence, b) Another possibility is to manually include the `domain name` into manually defined 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.

Contributions

The present study investigates the following objectives:

- 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, VisDA, and mini-DomainNet, we demonstrate the superior performance of AD-CLIP over state-of-the-art alternatives.

Architecture of AD-CLIP

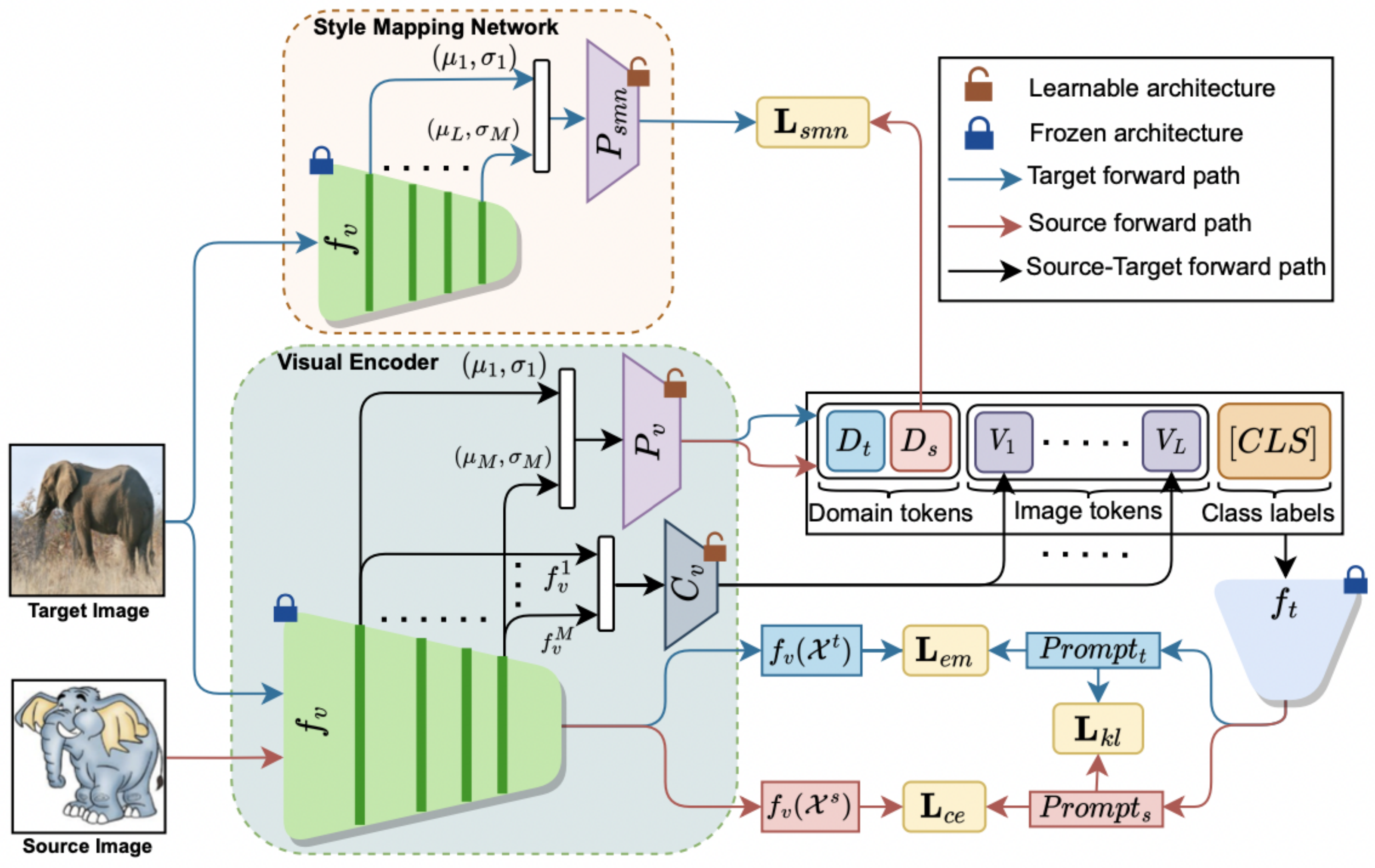


Figure 2. The architecture of AD-CLIP is based on the frozen CLIP backbones f_v and f_t. For prompt token learning, we introduce the new vision and text projectors P_v and C_v, respectively, which encoder the style and content information from the different layers of f_v. The style mapping network, P_{smn}, approximates the source domain style information from the target domain features. Best viewed in color.

Formulation of Metric Objectives

- Cross entropy loss:

$$\mathbf{L}_{ce} = CE(p(y|x), y) \tag{1}$$

where, the prediction probability of x for label y is defined as,

$$p(y|x) = \frac{\exp(\text{sim}(f_v(x), f_t(\text{Prompt}_y(x)))/\tau)}{\sum_{k=1}^{|\mathcal{Y}|} \exp(\text{sim}(f_v(x), f_t(\text{Prompt}_{y_k}(x)))/\tau)} \tag{2}$$

- Cross-domain style mapping loss:

$$\mathbf{L}_{smn} = \underset{P_{smn}, P_v}{\arg\min} \underset{P_{data}^{\mathcal{S}_l}, P_{data}^{\mathcal{T}_u}}{\mathbb{E}} ||D_s - P_{smn}(\bar{\mathcal{F}}_t)||_2^2 \tag{3}$$

- Domain alignment loss:

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

- Total loss: $\mathbf{L}_{total} = [\mathbf{L}_{ce} + \mathbf{L}_{smn} + \mathbf{L}_{Align}]$

Results and discussion

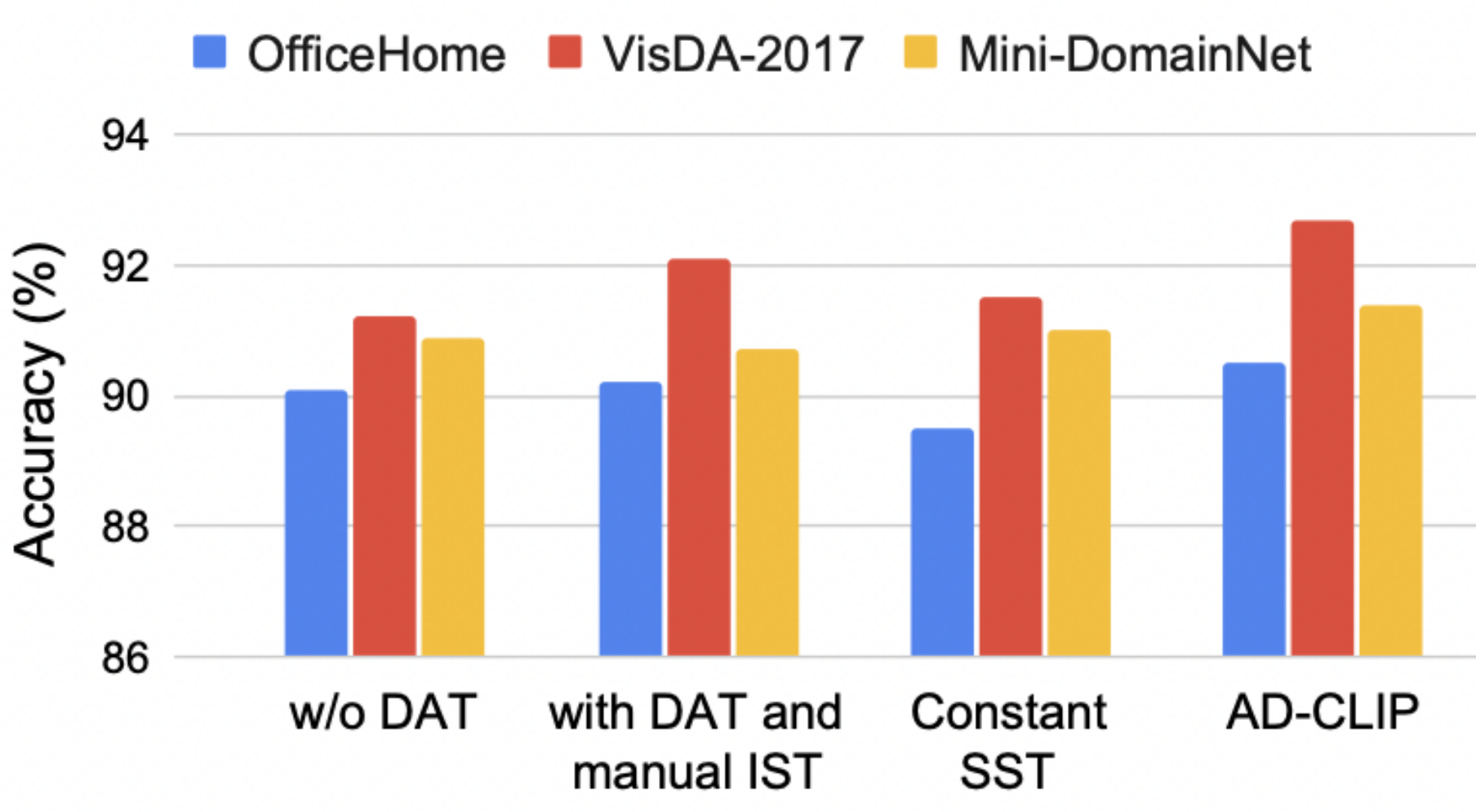
A. Unsupervised Domain Adaptation

Table 1. Comparison of AD-CLIP with state-of-the-art methods for UDA task on Office-Home dataset.

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	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		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		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		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		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		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		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		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		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** ± 0.1
CDTrans*	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		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		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		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		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		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** ± 0.2
CLIP	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		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		**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** ± 0.2

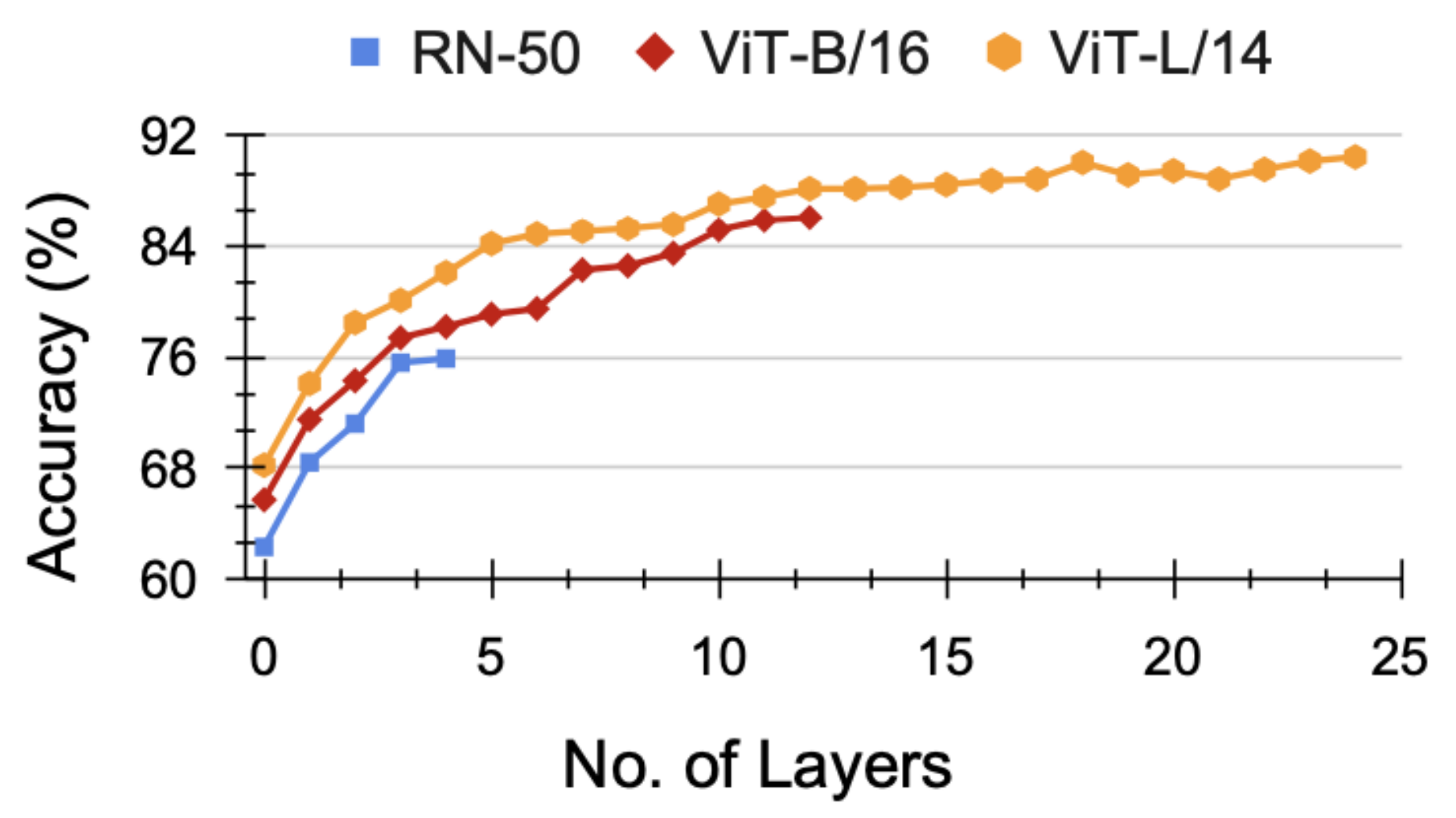
B. Sensitivity on prompt behaviour

Figure 3. DAT, IST and SST refer to domain-agnostic token, image-specific tokens and source- domain style tokens.



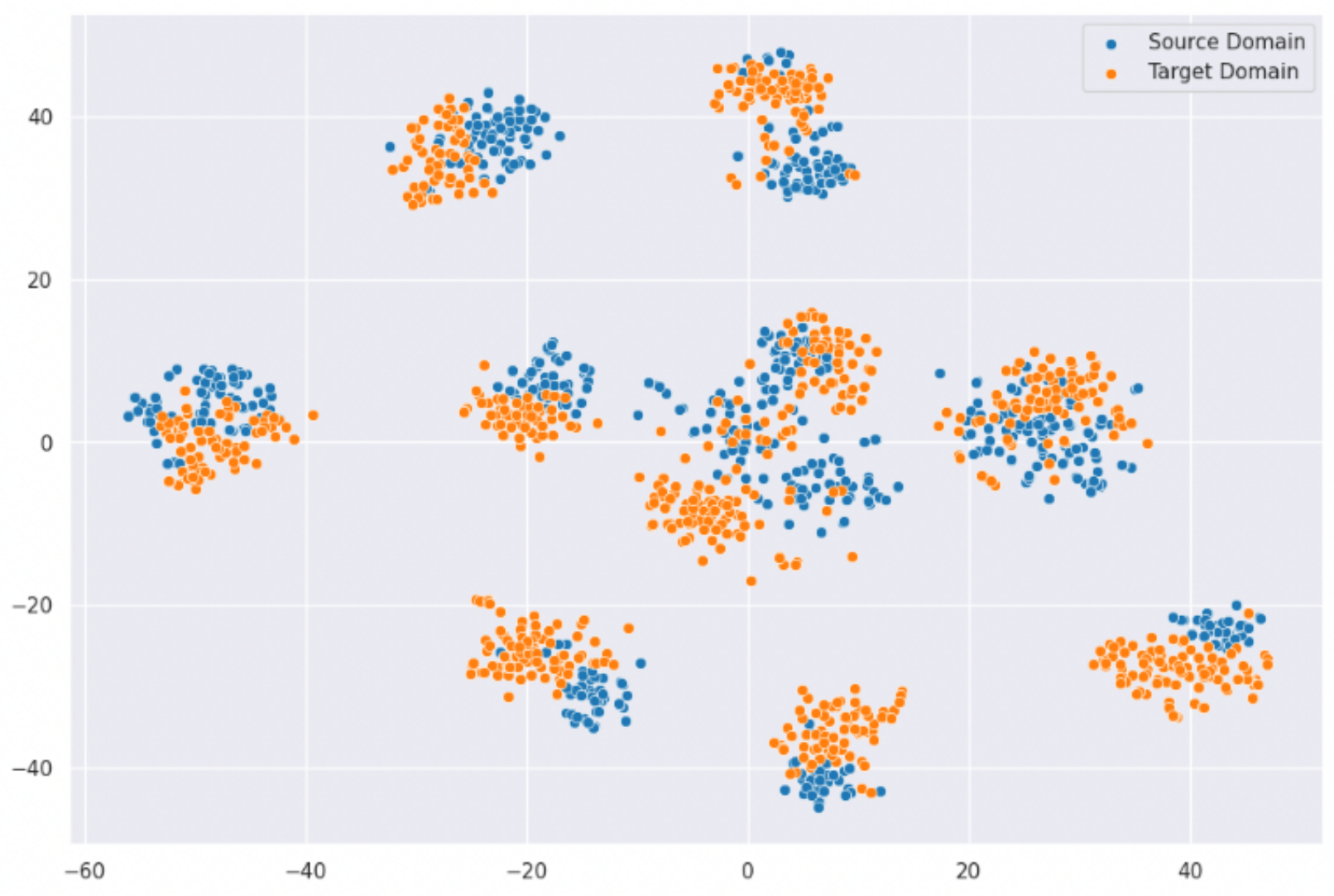
C. Sensitivity to the multi-scale features

Figure 4. 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.



D. t-SNE visualization

Figure 5. t-SNE of text embeddings from art and clipart domains of 10 classes of Office-Home.



Conclusions

- Introduced a domain adaptive model 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.
- 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, and learn three types of tokens in the prompts per image: domain token, image token, and class token.
- In the future, we plan to extend our approach to solve specific applications such as person re-identification and medical imaging.

References

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

[2] 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.

Authors

Mainak Singha

Harsh Pal

Ankit Jha

Biplab Banerjee