Towards Novel Domain Adaptation Techniques for Vision Tasks

A Dissertation
Submitted in partial fulfillment of
the requirements for the degree of
Bachelor of Technology in Electrical Engineering and Master of Technology in Communication and Signal Processing
by

Harsh Pal
(18D070012)

Supervisors:
Prof. Biplab Banerjee
and
Prof. Subhasis Chaudhuri

Department of Electrical Engineering
Indian Institute of Technology Bombay
Mumbai 400076 (India)

2023

Dissertation Approval

The dissertation entitled

Towards Novel Domain Adaptation Techniques for Vision Tasks

by

Harsh Pal (18D070012)

is approved for the degree of

Bachelor of Technology in Electrical Engineering and
Master of Technology in Communication & Signal Processing

Digital Signature
Biplab Banerjee (10001772)
26-Jun-23 08:15:05 PM

Prof. Biplab Banerjee
Department of Centre for Studies in Resource Engineering
(Supervisor)

Digital Signature
Abir De (10001916)
27-Jun-23 12:08:01 AM

Prof. Abir De
Department of Computer Science and Engineering
(Examiner)

Prof. Avik Hati
(Examiner)

Date: June 2023
Place: Mumbai.

Declaration

I declare that this written submission represents my ideas in my own words and where others' ideas or words have been included, I have adequately cited and referenced the original sources. I declare that I have properly and accurately acknowledged all sources used in the production of this report. I also declare that I have adhered to all principles of academic honesty and integrity and have not misrepresented or fabricated or falsified any idea/data/fact/source in my submission. I understand that any violation of the above will be a cause for disciplinary action by the Institute and can also evoke penal action from the sources which have thus not been properly cited or from whom proper permission has not been taken when needed.

Harsh Pal

Date: 1 July 2023 (18D070012)

Abstract

Deep Learning models have found significant utility in complex assignments like machine translation and autonomous driving. Nonetheless, constructing these models has proven to be difficult, primarily due to computational constraints and the need for abundant annotated data. Furthermore, these models may not perform well when faced with unfamiliar scenarios or data patterns in the target domain. Examples of such situations include transitioning between cities, varying weather conditions, or changes in lighting conditions. To address this issue, Unsupervised Domain Adaptation (UDA) leverages unlabeled data, which is readily available, to adjust models and make them more suitable for new conditions or data distributions. This thesis is based on the applications of two vision tasks, Image-based Shape retrieval (IBSR) and Image recognition. In IBSR, we discuss the transfer learning of shape retrieval of one dataset into another dataset which is a novel application in itself. In Image recognition, we discuss the application of large-scale foundation models like CLIP and prompt learning in UDA tasks.

Table of Contents

Abstract **v**

1 Introduction **1**

1.1 Background 2

1.1.1 Image Based shape retrieval 2

1.1.2 Unsupervised Domain Adaptation 2

1.1.3 Metric Learning 3

1.1.4 Prompt Learning 3

1.2 Structure of the Thesis 3

2 Related Works **5**

2.1 3D Shape Retrieval 5

2.2 Unsupervised Domain Adaptation 6

2.3 Metric Learning 7

3 Method **9**

3.1 Problem Definition 9

3.2 Overview of DA-IBSR 9

3.2.1 Training and Inference 10

4 Results and Discussions **11**

4.1 Dataset 11

4.2 Experimental Results 12

4.3 Ablations 13

4.4 Conclusions and Future Scope 14

5 Related Works **17**

5.1 Unsupervised Domain Adaptation 17

5.2 Prompt learning 17

6 Method **19**
6.1 Problem Definition 19
6.2 Overview of AD-CLIP 19
6.2.1 Prompt learning in AD-CLIP 20
6.2.2 Loss function for Domain Alignment 21

7 Results and Discussions **23**
7.1 Dataset 23
7.2 Experimental Results 23
7.2.1 Office-Home 23
7.2.2 VisDA-2017 24
7.2.3 Mini-Domainet 24
7.3 Ablation 25
7.4 Conclusion 26

References **27**

Chapter 1

Introduction

Deep Learning has gained immense popularity in recent years due to its impressive performance in numerous computer vision tasks and its wide range of applications. However, deep networks are comprised of a vast number of parameters, which poses computational challenges and hinders human comprehension of the network's decision-making process. For critical applications like histopathology, it is crucial for the model to understand and appropriately respond to various spatial changes. While networks are trained using data from a specific source domain, they may encounter a decline in performance when applied to a different target domain during inference (deployment). To ensure stable behavior, Domain Adaptation (DA) [Baktashmotlagh *et al.* (2013); Daumé III *et al.* (2010); Yao *et al.* (2015)] aims to adapt networks to new environments. Rather than building general models that should work in any given situation, DA focuses on agile approaches that utilize data from a source domain to develop models with high performance on a target domain. Unsupervised Domain Adaptation (UDA) takes this concept further by eliminating the need for annotations from the target domain. Instead, UDA achieves adaptation by leveraging unlabeled data from the target domain.

In practical scenarios, the data used to train a system often differs from the real-world environment where the system will be deployed. Additionally, it is usually not feasible to have labeled data that matches the exact distribution of the deployment environment. This has led to the rise of domain adaptation techniques, which are particularly popular in tasks where labeling is costly, and the quality of samples depends heavily on factors like camera setup and viewing angle.

UDA has seen substantial growth over the years, but it is still limited to distinct applications. So in this work, in the first part, we have introduced a novel application of UDA in Image-based shape retrieval. In this part, we have discussed the problem of domain shift not only in the conventional 2D space but also in the multi-view feature representation of 3D models. In the second part, we have discussed the application of DA in large-scale vision-based foundation models Jia *et al.* (2021) like CLIP Radford *et al.* (2021) for image recognition. We proposed a novel prompt learning paradigm by including vision features in the prompt.

1.1 Background

1.1.1 Image Based shape retrieval

3D model retrieval is a topic that has attracted much attention in 3D-related research, aiming to find other 3D models matching the query from the database. 3D models can be represented and stored in various forms, such as meshesFeng *et al.* (2018b,a), voxel gridsYang *et al.* (2022); Maturana and Scherer (2015), point cloudsYang *et al.* (2021) and 2D multiple viewsGao *et al.* (2020). When the query is in the form of data constructed directly on the original 3D model, such as voxels, point clouds, and multiple views, the retrieval task is only intra-modal and called model-based 3D model retrieval. When the query is a heterogeneous kind of data, such as an image, the retrieval task is designed to be a cross-domain problem.

1.1.2 Unsupervised Domain Adaptation

Unsupervised domain adaptation involves adapting a machine learning model from a source domain to a target domain, even when the data distributions differ. There are various approaches in the literature, such as distribution alignment, pseudo-labeling, and adversarial techniques. For instance, methods like Maximum Mean Discrepancy (MMD)Li *et al.* (2021) reduce the distribution distance in the kernel space, while DANNGanin and Lempitsky (2015) incorporates a domain classifier in a deep neural network to distinguish between source and target domains. Other approaches like CyCADAHoffman *et al.* (2017) and CDTransXu *et al.* (2021) utilize adversarial learning and attention mechanisms for feature alignment. Although vision-language models have been considered, the existing method DAPLGe *et al.* (2022) relies on ad-hoc prompting and manual inclusion of domain information.

1.1.3 Metric Learning

Metric learningKaya and Bilge (2019) aims to measure the similarity among samples while using an optimal distance metric for learning tasks. Metric learning methods, which generally use a linear projection, are limited in solving real-world problems demonstrating non-linear characteristics. Kernel approaches are utilized in metric learning to address this problem. In recent years, deep metric learning, which provides a better solution for nonlinear data through activation functions, has attracted researchers' attention in many different areas. SiameseWang *et al.* (2015) and Triplet networksGe (2018) are commonly used to correlate among samples while using shared weights in deep metric learning. The success of these networks is based on their capacity to understand the similarity relationship among samples.

1.1.4 Prompt Learning

Prompt learning, specifically in the context of large language models, refers to the process of training and fine-tuning these models using prompts or instructions to guide their behavior. Large language models, such as GPT-3Brown *et al.* (2020), have a vast amount of knowledge and can generate text in a wide range of contexts. However, prompt learning allows for more controlled and directed generation by explicitly providing instructions or prompts as input to the model. These prompts can take the form of a natural language sentence, a question, or a partial text. By carefully crafting and providing prompts, it is possible to influence the model's behavior, steer its responses, or prompt it to perform specific tasks. Zhou *et al.* (2022) optimize continuous vectors in the word embedding space.

1.2 Structure of the Thesis

This thesis is structured into two sections. The first part focuses on domain adaptation (DA) in IBSR or *DA-IBSR*, while the second part explores DA in large-scale vision-based models, specifically introducing *AD-CLIP*. Each part has 3 sections, Related Works, Method and Results & Discussions and each section has their subsections.

Part 1: DA-IBSR: Domain Adaptive Image Based Shape Retrieval

Chapter 2

Related Works

2.1 3D Shape Retrieval

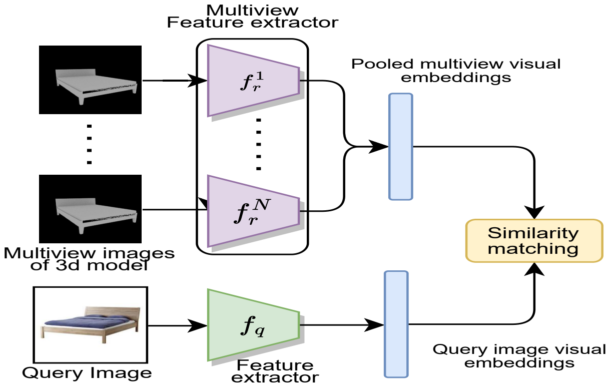


Figure 2.1: Architecture for Images based shape retrieval

Shape retrieval is one of the most fundamental problems in computer vision. With recent development in deep learning techniques for feature extraction and 3D shape datasets, 3D shape retrieval from single images based shape retrieval(IBSR) has gained more attention. Mu et al. Wu *et al.* (2018) proposed a novel architecture that maps two kinds of features into high-dimensional Hilbert space to decrease the gap. Deep cross-modality adaptation (DCA) employs a metric learning-based method to learn domain discriminative features and cross-modal transformation network to transfer the features of the 2D sketch to the 3D shape feature space. Zhou et al. (2019) proposed the unsupervised dual-level embedding alignment (DLEA) network, which was a first end-to-end network for this task. The gap between the two modalities is reduced by alignment at the embedding on domain and class levels. Recently, CDA (Hu, Zhou, Liu, et al., 2022) introduces a joint domain-class alignment module to learn a class-discriminative and domain-agnostic feature space for 2D images and 3D models. TDRL (Zhou et al., 2022) proposes to learn discriminative and transferable cross-domain representation for 2D and 3D data using unsupervised adversarial domain adaptation.

Despite significant prior works, using single images to retrieve the 3D shapes is still a challenging problem, and the major reason for this challenge is the problem of domain shift in both 2D and 3D features space. Solution to this problem is Domain Adaptation(DA) Baktashmotlagh *et al.* (2013); Daumé III *et al.* (2010); Yao *et al.* (2015). For our work, we will be focusing on Unsupervised Domain adaptation techniques.

2.2 Unsupervised Domain Adaptation

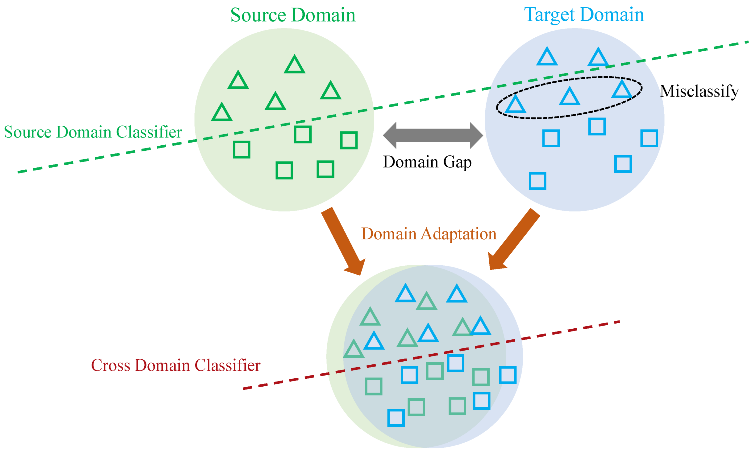


Figure 2.2: Unsupervised Domain Adaptation

DA refers to the process of adjusting a machine learning model that has been trained on one dataset (the source domain) to perform well on a different dataset (the target domain) where the data distributions may vary. The existing literature offers a wide range of approaches for domain adaptation, including techniques such as sub-space alignment, pseudo-labeling, and adversarial methods, among others. For instance, Maximum Mean Discrepancy (MMD)**?** is a popular technique that aims to minimize the discrepancy between the distributions of the source and target domains in the kernel space. Another commonly used approach is Domain-Adversarial Neural Networks (DANN)Ganin and Lempitsky (2015), which involves incorporating a domain classifier into the deep neural network to enable it to differentiate between source and target domain data. CDTransXu *et al.* (2021) employs cross-attention and two-way center-aware labeling in Transformers**?** to achieve domain alignment, making it robust against noisy label pairs. Recently, there has been interest in exploring vision-language models to tackle the domain adaptation task, given their improved feature space. The current method in this domain, DAPLGe *et al.* (2022), relies on ad-hoc prompting to learn disentangled domain and category representations.

2.3 Metric Learning

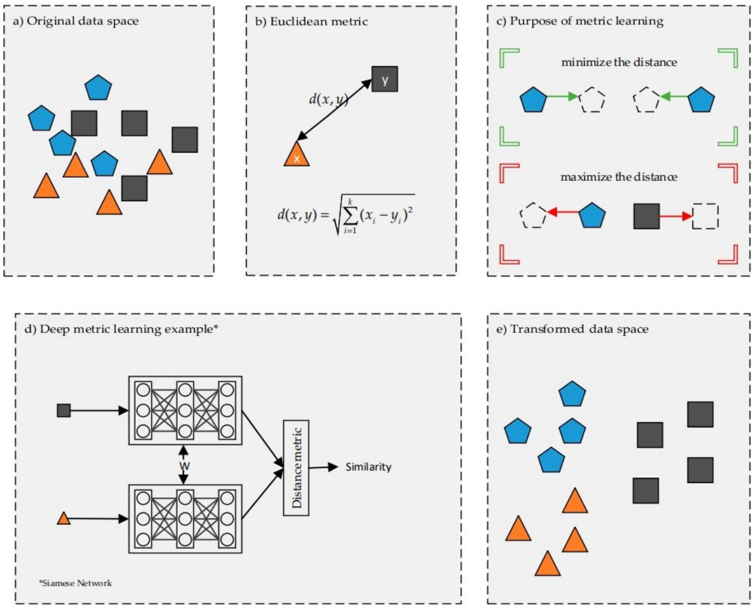


Figure 2.3: Metric Learning

Traditional machine learning techniques are constrained in their ability to handle raw data effectively. Hence, they require feature engineering, involving preprocessing and feature extraction steps, prior to performing classification or clustering tasks. These steps demand expertise and operate separately from the classification process. Conversely, deep learning directly learns high-level representations of data within the classification structure.

Common similarity metrics used for data classification include Euclidean, Mahalanobis, Matusita [Matusita (1955)], Bhattacharyya [Aherne *et al.* (1998)], and Kullback-Leibler distances [Elgammal *et al.* (2003)]. However, these predefined metrics have limited capabilities when it comes to data classification. Typically, metric learning approaches involve linear transformations of data without incorporating kernel functions, which are inadequate for capturing the nonlinear characteristics [Yu *et al.* (2016)] of the data. Deep learning overcomes this limitation by employing activation functions with nonlinear structures.

Existing deep learning approaches mostly focus on the deep architecture itself rather than the distance metric in a newly constructed representation space of the data. However, distance-based approaches have recently gained significant attention in the realm of deep learning [Duan *et al.* (2017), Dai *et al.* (2018)]. Deep metric learning, which aims to decrease the distance between dissimilar samples and increase the distance between similar samples, has emerged as an intriguing topic within this domain

Chapter 3

Method

3.1 Problem Definition

Given the image samples $\mathbf{s}_I^j$ and corresponding class labels $y_{\mathcal{S}}^j$, our objective is to retrieve the 3D shapes from the provided dataset. The scope of this research extends to a cross-domain setting, wherein the retrieval model is primarily trained on the data originating from the source domain $\mathcal{S}$ and subsequently deployed in another domain denoted as $\mathcal{T}$. To address the challenge of domain discrepancy, we adopt a transductive learning approach, leveraging unpaired and unlabeled query 2D images and 3D shapes from the target domain during the training phase to facilitate domain alignment.

3.2 Overview of DA-IBSR

Schematic of proposed DA-IBSR from cross-domain 3D shape retrieval. Our model addresses the different aspects of modality alignment and domain alignment in transductive setting to perform shape retrieval from target domain, when trained with labels and pairs from source domain and unpaired and unlabelled samples from target domain. Initially, we pass the 2D images (from $\mathcal{S}$ and $\mathcal{T}$) through the feature extractors $\mathcal{G}_q$, thus getting embedding $(f_q)_{\mathcal{S}}$ and $(f_q)_{\mathcal{T}}$. Simultaneously, 3D shapes are passed through $\mathcal{G}_r$ to get embeddings $(f_r)_{\mathcal{S}}$, $(f_r)_{\mathcal{T}}$ and $(f_r)_{\mathcal{S}}^-$ (the last one being the embedding of the negative sample). For modality alignment, we primarily rely on the triplet loss, which minimizes the Euclidean distance between $(f_q)_{\mathcal{S}}$ and $(f_r)_{\mathcal{S}}$, while maximizes the same between $(f_q)_{\mathcal{S}}$ and $(f_r)_{\mathcal{S}}^-$. Additionally, the features $(f_q)_{\mathcal{S}}$ and $(f_r)_{\mathcal{S}}$ are passed through the classifiers C_q and C_r to get the corresponding probabilities, which are compared against the available groundtruth. For domain alignment, we minimize the maximum mean discrepancy loss between the 2D ($(f_q)_{\mathcal{S}}$ and $(f_q)_{\mathcal{T}}$) and 3D ($(f_r)_{\mathcal{S}}$ and $(f_r)_{\mathcal{T}}$) embeddings of the two domains

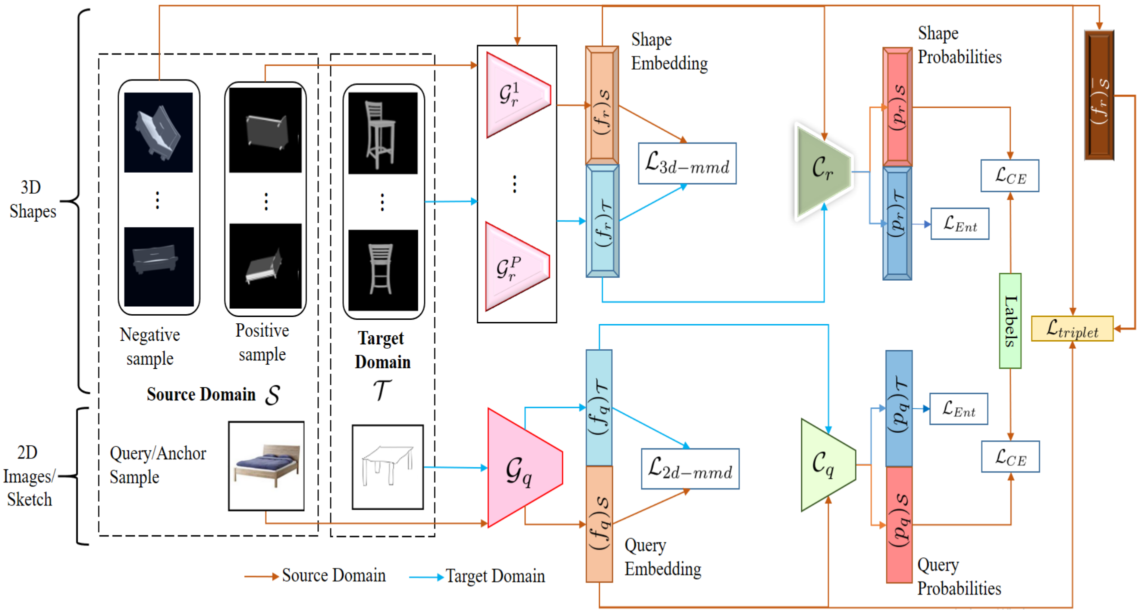


Figure 3.1: Proposed Architecture for **DA-IBSR**

each. Furthermore, for better intraclass discriminativeness in $\mathcal{T}$, we employ an entropy loss for both the modalities.

3.2.1 Training and Inference

The model is trained on the combined loss function given as:

$$\begin{aligned}\mathcal{L}_{final} = (\lambda_{2D}\mathcal{L}_{mmd_{2D}} + \lambda_{3D}\mathcal{L}_{mmd_{3D}}) + (\lambda_{c_q}\mathcal{L}_{c_q} + \\ \lambda_{c_r}\mathcal{L}_{c_r}) + (\lambda_{triplet}\mathcal{L}_{triplet}) + (\lambda_{E_q}\mathcal{L}_{E_q} + \\ \lambda_{E_r}\mathcal{L}_{E_r}) \qquad (3.1)\end{aligned}$$

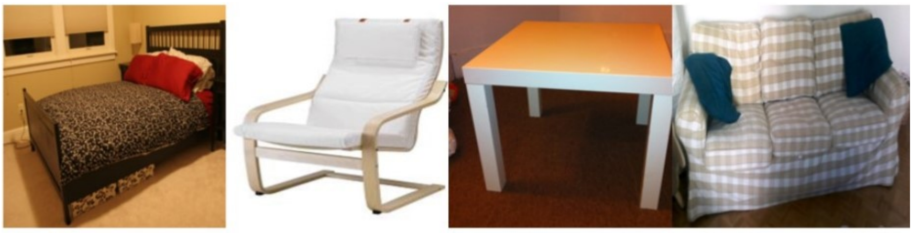
Here, all the λ terms represent the Lagrangian multipliers for the corresponding loss functions.

After the training is complete, we take the corresponding feature extractors for the 2D and 3D modalities from the target set $\mathcal{T}$, and pass the query images and shapes from the retrieval dataset through them. Then the similarity between the embedding of query and the shapes is calculated using L2 norm, and the images are retrieved in the ascending order of the magnitude of the L2-norm. In order to assess the performance of shape retrieval, we use mean average precision (mAP) Kishida (2005), calculated over K number of retrieved samples.

Chapter 4

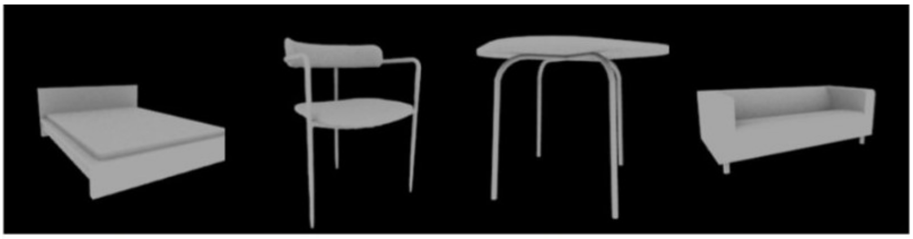
Results and Discussions

4.1 Dataset



(a) Bed (b) Chair (c) Table (d) Sofa

Figure 4.1: Sample Pix3D Query images



(a) Bed (b) Chair (c) Table (d) Sofa

Figure 4.2: Sample Pix3D Rendering images

We used three different benchmark datasets for our experimentation, SHREC'14Godil and Li (2014), Pix3DSun *et al.* (2018) and ShapeNetChang *et al.* (2015). SHREC'14 contains 13,680 hand-drawn sketches and 8,987 3D shapes. This dataset also contains 80 sketches for each category, 50 for training and 30 for testing. We combined both the split into one set. Pix3D consists of 10069 real-world images and 395 unique 3D models. ShapeNet consists of 51,300 3D shape models, the query images for ShapeNet were taken from ImageNetDeng *et al.* (2009), which consists of 1.4 million images across 1000 categories. We chose 4 common classes across these 4 datasets for our experiments, viz. bed, chair, sofa and table.

4.2 Experimental Results

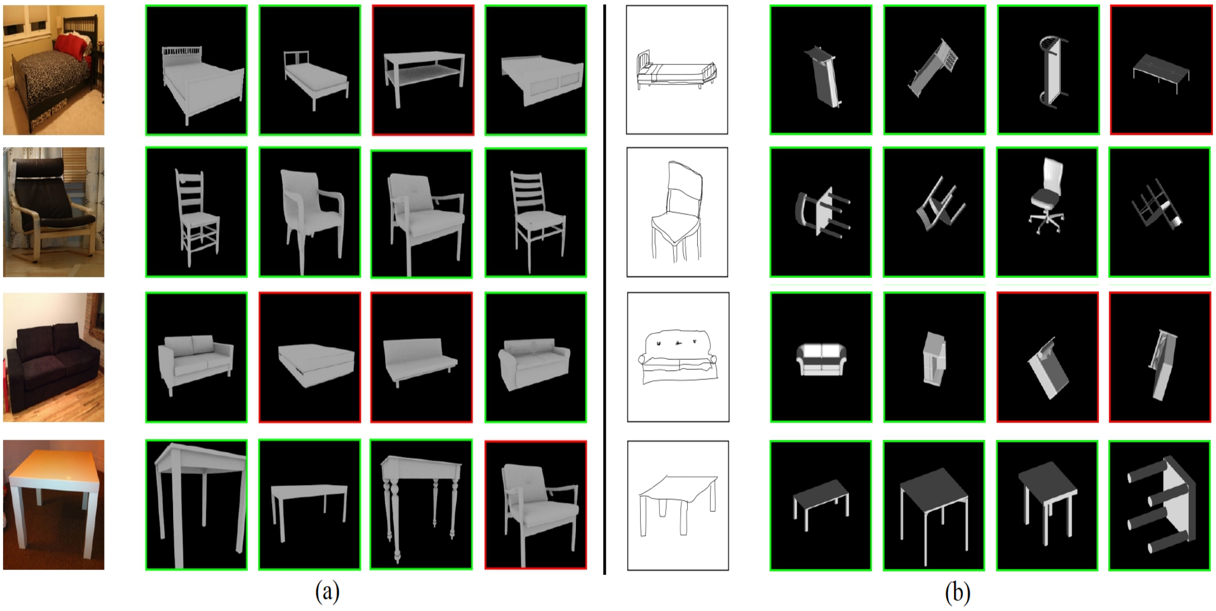


Figure 4.3: Cross domain retrieval of 3D shapes from (a) RGB images and (b) Sketches

Table 4.1: mAP analysis using our proposed approach on SHREC'14 and Pix3D datasets. In the experiments, we have alternately used SHREC'14 and Pix3D as source and target domains.

Method	SHREC→ Pix3D	Pix3D→ SHREC
IBSR **?**	0.14	0.24
DD-GAN **?**	0.42	0.25
DA - MSE Loss (ResNet18)	0.34	0.22
DA - Triplet Loss (ResNet18)	**0.70**	**0.36**
DA - MSE Loss (ResNet134)	0.45	0.30
DA - Triplet Loss (ResNet34)	**0.75**	**0.42**

The results of our proposed approach are presented in Tables 4.1 and 4.2 respectively for SHREC'14/Pix3D and SHREC'14/ShapeNet datasets. The first two rows show the results of cross-domain retrieval using methods that were trained on single domain setting (IBSR **?** and DD-GAN **?**). We find that DD-GAN outperforms IBSR owing to the effort to generate modality invariant features by the former, while IBSR only focuses on color

Table 4.2: mAP analysis using our proposed approach on SHREC’14 and ShapeNet datasets. In the experiments, we have alternately used SHREC’14 and ShapeNet as source and target domains.

Method	SHREC → ShapeNet	ShapeNet → SHREC
IBSR **?**	0.27	0.24
DD-GAN **?**	0.40	0.27
DA - MSE Loss (ResNet18)	0.28	0.30
DA - Triplet Loss (ResNet18)	**0.62**	**0.37**
DA - MSE Loss (ResNet34)	0.30	0.33
DA - Triplet Loss (ResNet34)	**0.64**	**0.41**

invariant features. However, when compared against DA-IBSR, it is visible that the mAP is much better in comparison to when DA is not applied. In addition, to highlight the effectiveness of triplet loss for modality alignment, we have compared it against mean squared error loss (MSE). It is clearly visible that triplet loss outperforms the MSE loss by a large margin, in both the dataset pairs and for both ResNet18 and ResNet34 feature extractors. In addition, it is also visible in both the tables that when SHREC dataset (consisting of sketches) is chosen as source domain, the retrieval performance on Pix3D and ShapeNet (which are image based datasets) is much better in contarst to the case when Pix3D or ShapeNet chosen as source domain and test on SHREC

4.3 Ablations

Table 4.3: Ablation study over the different combination of domain adaptive and modality alignment losses in DAIS-Net on SHREC’14→Pix3D dataset pair.

Loss combination	mAP
$\mathcal{L}_{mmd_{2D}} + \mathcal{L}_{mmd_{3D}} + \mathcal{L}_{CE}$	0.47
$\mathcal{L}_{mmd_{3D}} + \mathcal{L}_{triplet} + \mathcal{L}_{CE}$	0.48
$\mathcal{L}_{mmd_{2D}} + \mathcal{L}_{triplet} + \mathcal{L}_{CE}$	0.57
$\mathcal{L}_{mmd_{2D}} + \mathcal{L}_{mmd_{3D}} + \mathcal{L}_{triplet}$	0.66
$\mathcal{L}_{mmd_{2D}} + \mathcal{L}_{mmd_{3D}} + \mathcal{L}_{triplet} + \mathcal{L}_{CE}$	0.71
$\mathcal{L}_{mmd_{2D}} + \mathcal{L}_{mmd_{3D}} + \mathcal{L}_{triplet} + \mathcal{L}_{CE} + \mathcal{L}_{E}$	**0.75**

We have performed an ablation study on the losses, the ablation study of which is shown in Table 4.3 for SHREC’14 and Pix3D dataset (with the former one being the source domain) with ResNet34 as the backbone model. For simplicity, the entropy and classification losses for both the modalities are represented as $\mathcal{L}_{CE}$ and $\mathcal{L}_{E}$. It is clearly visible that when all the losses are involved, mAP is the highest (0.75), thus proving that all the losses complement each other and contribute together in learning a shared cross-modal and cross-domain representation

4.4 Conclusions and Future Scope

In this work, we propose a pioneering approach of multimodal domain adaptation for image-based 3D shape retrieval. Our method tackles the challenges of aligning two modalities (2D images/sketches and 3D shapes) and aligning disjoint domains containing the 2D queries and 3D shapes. To address the modality alignment challenge, we introduce the notion of negative sample mining and in the source domain and create the negative samples from 3D shapes from the classes different from that of the query image. Then, we employ the triplet loss where the distance between the samples of same class is minimised and that of two different classes is maximised between 2D image and 3D shape. Furthermore, we incorporate modality-specific classifiers to align the 3D shapes and 2D images in the source domain. To address domain alignment, we utilize the maximum mean discrepancy for each modality across the two domains. Additionally, we introduce an entropy loss to establish discriminativeness among the target features based on class probabilities of the target embeddings for both modalities. Our model is evaluated on two pairs of image-shape datasets, namely SHREC'14-Pix3D and Shrec'14-ShapeNet. The experimental results demonstrate the presence of the domain adaptation problem, and our proposed approach effectively addresses it, as evidenced by both qualitative and quantitative evaluations. As future work, we plan to extend our approach to zero-shot domain adaptation and explore domain generalization techniques to handle multiple domains simultaneously.

Part 2: AD-CLIP: Adapting Domains in Prompt Space Using CLIP

Chapter 5

Related Works

5.1 Unsupervised Domain Adaptation

DA refers to the process of adjusting a machine learning model that has been trained on one dataset (the source domain) to perform well on a different dataset (the target domain) where the data distributions may vary. The existing literature offers a wide range of approaches for domain adaptation, including techniques such as sub-space alignment, pseudo-labeling, and adversarial methods, among others. For instance, Maximum Mean Discrepancy (MMD)**?** is a popular technique that aims to minimize the discrepancy between the distributions of the source and target domains in the kernel space. Another commonly used approach is Domain-Adversarial Neural Networks (DANN)Ganin and Lempitsky (2015), which involves incorporating a domain classifier into the deep neural network to enable it to differentiate between source and target domain data. CDTransXu *et al.* (2021) employs cross-attention and two-way center-aware labeling in Transformers**?** to achieve domain alignment, making it robust against noisy label pairs. Recently, there has been interest in exploring vision-language models to tackle the domain adaptation task, given their improved feature space.

5.2 Prompt learning

The basic idea of knowledge probing is to induce pre-trained language models to generate answers given cloze-style prompts, which can benefit a number of downstream tasks, such as sentiment analysis. [Jiang *et al.* (2020)] propose to generate candidate prompts through text mining and paraphrasing, and identify the optimal ones that give the highest training accuracy. [Shin *et al.* (2020)] introduce a gradient based approach, which searches for tokens with the largest gradient changes in the label likelihood. Most related to our work are continuous prompt learning methods [Zhang *et al.* (2022), Zhong *et al.* (2021)] which optimize continuous vectors in the word embedding space. A drawback of such methods compared to searching discrete tokens is the lack of a clear way to visualize what “words” are learned for the vectors.

Chapter 6

Method

6.1 Problem Definition

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

6.2 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

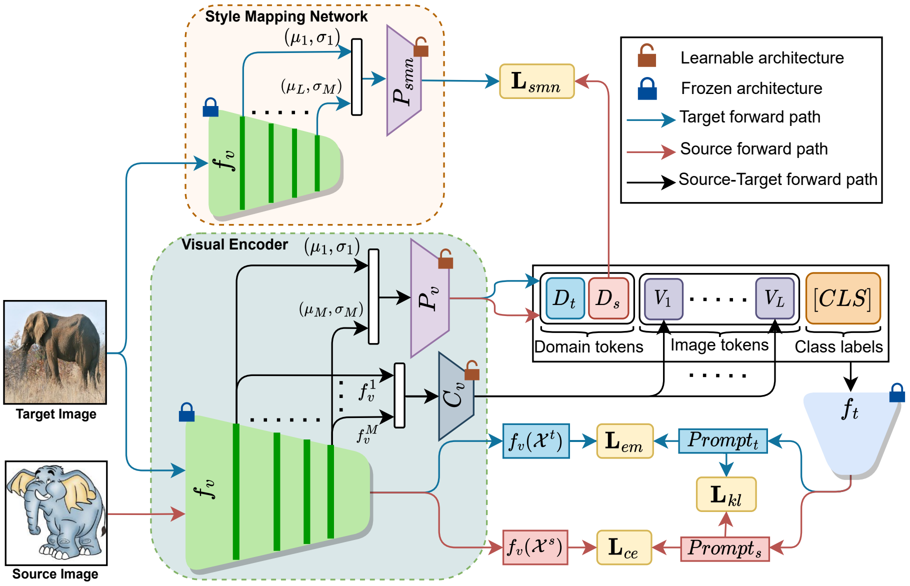


Figure 6.1: The architecture of AD-CLIP is based on the frozen CLIP backbones f_v and f_t . For prompt token learning, we introduce the new vision and text projectors P_v and C_v, respectively, which encoder the style and content information from the different layers of f_v. The style mapping network, Psmn, approximates the source domain style information from the target domain features. Best viewed in color

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 loss functions for classification and domain alignment.

6.2.1 Prompt learning in AD-CLIP

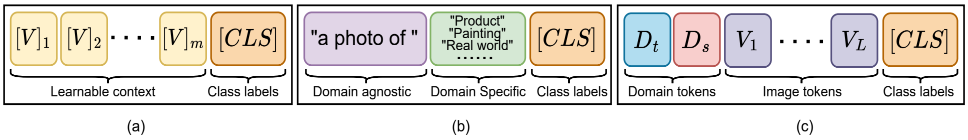


Figure 6.2: We highlight the differences between our prompts from the literature. a) CoOp [Radford *et al.* (2021)] directly learns the prompt tokens from random vectors and may not be suitable for DA as it does not concern domain independence, b) Another possibility is to manually include the domain 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.

Our objective is to learn prompts directly from the visual domain to effectively encode the visual distribution, as opposed to the static prompting technique **?**. 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 $\bar{\mathcal{F}}(x)$.

6.2.2 Loss function for Domain Alignment

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

$$\mathbf{L}_{Align} = \underset{P_v, C_v}{\arg\min} \underset{(x,y)\in P^{\mathcal{S}_l}_{data}}{\mathbb{E}} \mathbf{L}_{em}([p(y_1|x); p(y_2|x), \cdots; p(y_{|\mathcal{Y}|}|x)]) + \mathbf{L}_{KL}(\text{Prompt}_t|\text{Prompt}_s) \tag{6.1}$$

$$p(y|x) = \frac{\exp(\text{sim}(f_v(x), f_t(\text{Prompt}_y(x)))/\tau)}{\sum_{k=1}^{|\mathcal{Y}|} \exp(\text{sim}(f_v(x), f_t(\text{Prompt}_{y_k}(x)))/\tau)} \tag{6.2}$$

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

$$\mathbf{L}_{smn} = \underset{f_{smn}, P_v}{\arg\min} \underset{P^{\mathcal{S}_l}_{data}, P^{\mathcal{T}_u}_{data}}{\mathbb{E}} ||D_s - f_{smn}(\bar{\mathcal{F}}_t)||_2^2 \tag{6.3}$$

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

Chapter 7

Results and Discussions

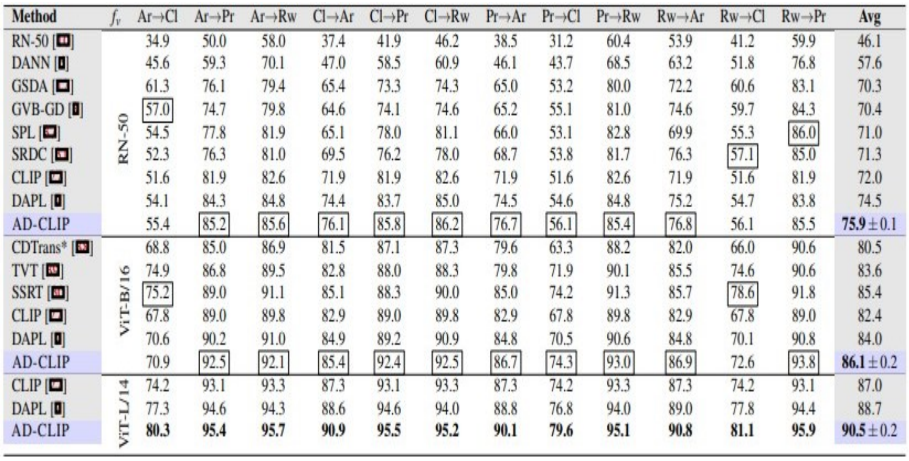
7.1 Dataset

We validate our model on three publicly available DA datasets. i) **Office-Home** Venkateswara *et al.* (2017): 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** Peng *et al.* (2017): 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 Long *et al.* (2018); Saito *et al.* (2018), 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 Peng *et al.* (2019) called Mini-DomainNet. This subset features four domains, including Clipart (c), Painting (p), Real (r), and Sketch (s), each with images from 126 categories.

7.2 Experimental Results

7.2.1 Office-Home

Comparison of AD-CLIP with state-of-the-art methods for UDA task on VisDA-2017 Peng *et al.* (2017) dataset. We show our results for every class with three different vision backbones. However, CDTrans* has used DeiT-base **?** backbone only. The overall best accuracy and best within per backbone are indicated in bold and box respectively.

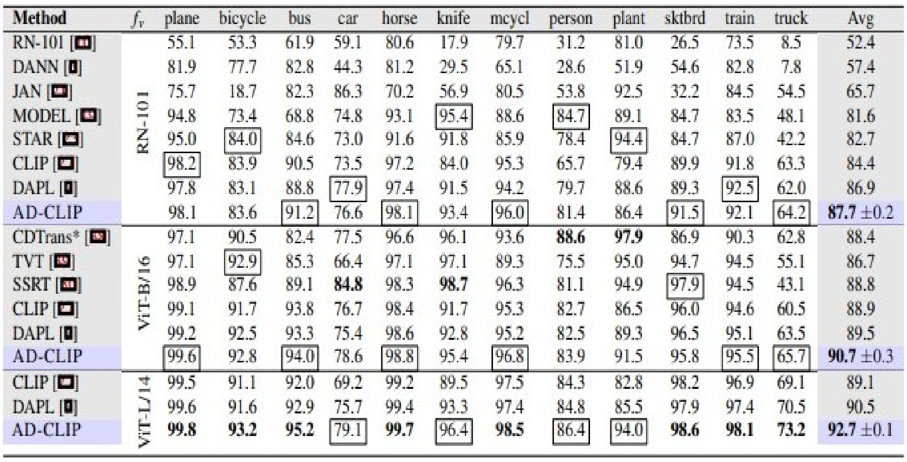


Method	f_v	Ar→Cl	Ar→Pr	Ar→Rw	Cl→Ar	Cl→Pr	Cl→Rw	Pr→Ar	Pr→Cl	Pr→Rw	Rw→Ar	Rw→Cl	Rw→Pr	Avg
RN-50 [■]	RN-50	34.9	50.0	58.0	37.4	41.9	46.2	38.5	31.2	60.4	53.9	41.2	59.9	46.1
DANN [■]		45.6	59.3	70.1	47.0	58.5	60.9	46.1	43.7	68.5	63.2	51.8	76.8	57.6
GSDA [■]		61.3	76.1	79.4	65.4	73.3	74.3	65.0	53.2	80.0	72.2	60.6	83.1	70.3
GVB-GD [■]		$\boxed{57.0}$	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	$\boxed{86.0}$	71.0
SRDC [■]		52.3	76.3	81.0	69.5	76.2	78.0	68.7	53.8	81.7	76.3	$\boxed{57.1}$	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	$\boxed{85.2}$	$\boxed{85.6}$	$\boxed{76.1}$	$\boxed{85.8}$	$\boxed{86.2}$	$\boxed{76.7}$	$\boxed{56.1}$	$\boxed{85.4}$	$\boxed{76.8}$	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 [■]		$\boxed{75.2}$	89.0	91.1	85.1	88.3	90.0	85.0	74.2	91.3	85.7	$\boxed{78.6}$	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	$\boxed{92.5}$	$\boxed{92.1}$	$\boxed{85.4}$	$\boxed{92.4}$	$\boxed{92.5}$	$\boxed{86.7}$	$\boxed{74.3}$	$\boxed{93.0}$	$\boxed{86.9}$	72.6	$\boxed{93.8}$	**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

Figure 7.1: Comparison for Office-Home

7.2.2 VisDA-2017

Comparison of AD-CLIP with the state-of-the-art vision-language models for UDA task on Mini-DomainNet Peng *et al.* (2019) dataset. The overall best accuracy and best within per backbone are indicated in bold and box respectively.

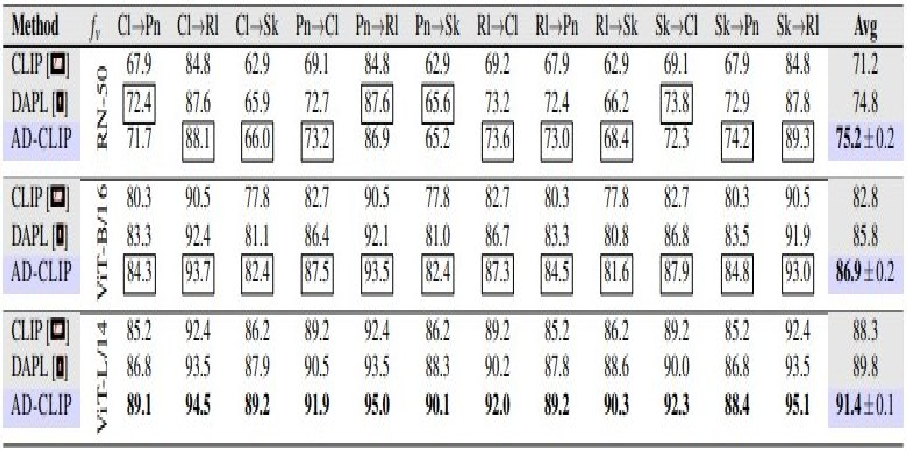


Method	f_v	plane	bicycle	bus	car	horse	knife	mcycl	person	plant	sktbrd	train	truck	Avg
RN-101 [■]	RN-101	55.1	53.3	61.9	59.1	80.6	17.9	79.7	31.2	81.0	26.5	73.5	8.5	52.4
DANN [■]		81.9	77.7	82.8	44.3	81.2	29.5	65.1	28.6	51.9	54.6	82.8	7.8	57.4
JAN [■]		75.7	18.7	82.3	86.3	70.2	56.9	80.5	53.8	92.5	32.2	84.5	54.5	65.7
MODEL [■]		94.8	73.4	68.8	74.8	93.1	$\boxed{95.4}$	88.6	$\boxed{84.7}$	89.1	84.7	83.5	48.1	81.6
STAR [■]		95.0	$\boxed{84.0}$	84.6	73.0	91.6	91.8	85.9	78.4	$\boxed{94.4}$	84.7	87.0	42.2	82.7
CLIP [■]		$\boxed{98.2}$	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 [■]		97.8	83.1	88.8	$\boxed{77.9}$	97.4	91.5	94.2	79.7	88.6	89.3	$\boxed{92.5}$	62.0	86.9
AD-CLIP		98.1	83.6	$\boxed{91.2}$	76.6	$\boxed{98.1}$	93.4	$\boxed{96.0}$	81.4	86.4	$\boxed{91.5}$	92.1	$\boxed{64.2}$	**87.7** ±0.2
CDTrans* [■]	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 [■]		97.1	$\boxed{92.9}$	85.3	66.4	97.1	97.1	89.3	75.5	95.0	94.7	94.5	55.1	86.7
SSRT [■]		98.9	87.6	89.1	**84.8**	98.3	**98.7**	96.3	81.1	94.9	$\boxed{97.9}$	94.5	43.1	88.8
CLIP [■]		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 [■]		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		$\boxed{99.6}$	92.8	$\boxed{94.0}$	78.6	$\boxed{98.8}$	95.4	$\boxed{96.8}$	83.9	91.5	95.8	$\boxed{95.5}$	$\boxed{65.7}$	**90.7** ±0.3
CLIP [■]	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 [■]		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		**99.8**	**93.2**	**95.2**	$\boxed{79.1}$	**99.7**	$\boxed{96.4}$	**98.5**	$\boxed{86.4}$	$\boxed{94.0}$	**98.6**	**98.1**	**73.2**	**92.7** ±0.1

Figure 7.2: Comparison for VisDA-2017

7.2.3 Mini-Domainet

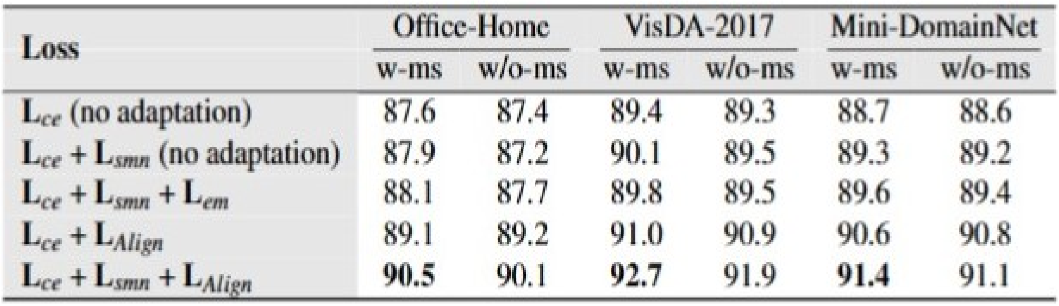
Comparison of AD-CLIP with the state-of-the-art vision-language models for UDA task on Mini-DomainNet Peng *et al.* (2019) dataset. The overall best accuracy and best within per backbone are indicated in bold and box respectively.



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 [■]	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 [■]		$\boxed{72.4}$	87.6	65.9	72.7	$\boxed{87.6}$	$\boxed{65.6}$	73.2	72.4	66.2	$\boxed{73.8}$	72.9	87.8	74.8
AD-CLIP		71.7	$\boxed{88.1}$	$\boxed{66.0}$	$\boxed{73.2}$	86.9	65.2	$\boxed{73.6}$	$\boxed{73.0}$	$\boxed{68.4}$	72.3	$\boxed{74.2}$	$\boxed{89.3}$	**75.2**±0.2
CLIP [■]	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 [■]		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		$\boxed{84.3}$	$\boxed{93.7}$	$\boxed{82.4}$	$\boxed{87.5}$	$\boxed{93.5}$	$\boxed{82.4}$	$\boxed{87.3}$	$\boxed{84.5}$	$\boxed{81.6}$	$\boxed{87.9}$	$\boxed{84.8}$	$\boxed{93.0}$	**86.9**±0.2
CLIP [■]	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 [■]		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**±0.1

Figure 7.3: Comparison for Mini-Domainet

7.3 Ablation



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

Figure 7.4: 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.

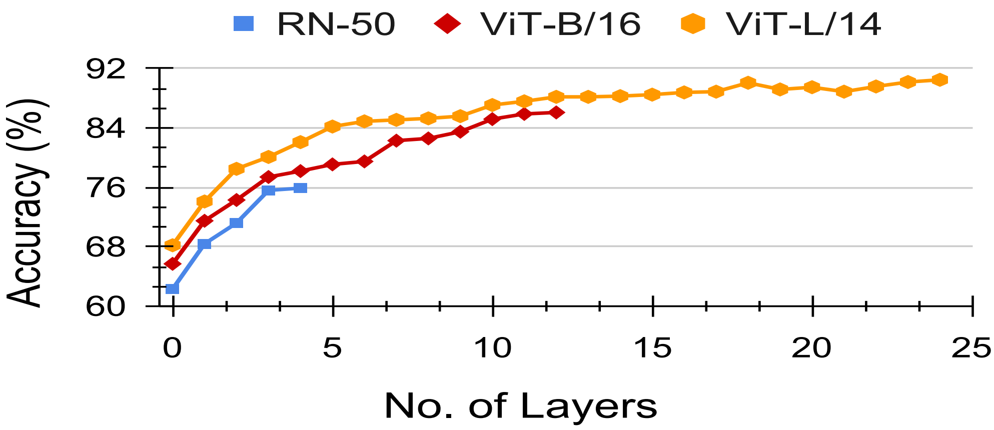


Figure 7.5: Performance of AD-CLIP with different layers of RN50, ViT- B/16 and ViT-L/14 backbones to extract multi-scale features on Office-Home.

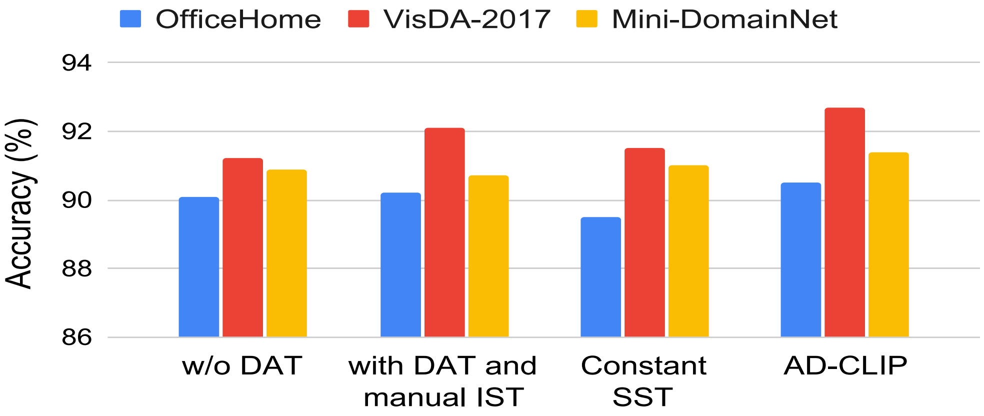


Figure 7.6: Compar- ison of results of AD-CLIP with different prompt settings. Here DAT, IST and SST refer to domain-agnostic token, image-specific tokens and source-domain style tokens

7.4 Conclusion

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.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

Aherne, F. J., Thacker, N. A., and Rockett, P. I., 1998, “The bhattacharyya metric as an absolute similarity measure for frequency coded data,” *Kybernetika* **34**, 363–368.

Baktashmotlagh, M., Harandi, M. T., Lovell, B. C., and Salzmann, M., 2013, “Unsupervised domain adaptation by domain invariant projection,” in *2013 IEEE International Conference on Computer Vision*, pp. 769–776.

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., and Amodei, D., 2020, “Language models are few-shot learners,”

Chang, A. X., Funkhouser, T., Guibas, L., Hanrahan, P., Huang, Q., Li, Z., Savarese, S., Savva, M., Song, S., Su, H., *et al.*, 2015, “Shapenet: An information-rich 3d model repository,” *arXiv preprint arXiv:1512.03012*

Dai, G., Xie, J., and Fang, Y., 2018, “Deep correlated holistic metric learning for sketch-based 3d shape retrieval,” *IEEE Transactions on Image Processing* **27**, 3374–3386.

Daumé III, H., Kumar, A., and Saha, A., 2010 Jul., “Frustratingly easy semi-supervised domain adaptation,” in *Proceedings of the 2010 Workshop on Domain Adaptation for Natural Language Processing* (Association for Computational Linguistics, Uppsala, Sweden). pp. 53–59.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L., 2009, “Imagenet: A large-scale hierarchical image database,” in *2009 IEEE conference on computer vision and pattern recognition* (Ieee). pp. 248–255.

Duan, Y., Lu, J., Feng, J., and Zhou, J., 2017, “Deep localized metric learning,” *IEEE Transactions on Circuits and Systems for Video Technology* **28**, 2644–2656.

Elgammal, A., Duraiswami, R., and Davis, L. S., 2003, “Probabilistic tracking in joint feature-spatial spaces,” in *2003 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2003. Proceedings.*, Vol. 1 (IEEE). pp. I–I.

Feng, Y., Zhang, Z., Zhao, X., Ji, R., and Gao, Y., 2018a, "Gvcnn: Group-view convolutional neural networks for 3d shape recognition," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 264–272.

Feng, Y., Feng, Y., You, H., Zhao, X., and Gao, Y., 2018b, "Meshnet: Mesh neural network for 3d shape representation,"

Ganin, Y., and Lempitsky, V., 2015, "Unsupervised domain adaptation by backpropagation," in *International conference on machine learning* (PMLR). pp. 1180–1189.

Gao, Z., Li, Y., and Wan, S., 2020 feb, "Exploring deep learning for view-based 3d model retrieval," *ACM Trans. Multimedia Comput. Commun. Appl.* **16**, doi:\bibinfo{doi}{10.1145/3377876}

Ge, C., Huang, R., Xie, M., Lai, Z., Song, S., Li, S., and Huang, G., 2022, "Domain adaptation via prompt learning," *arXiv preprint arXiv:2202.06687*

Ge, W., 2018, "Deep metric learning with hierarchical triplet loss," in *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 269–285.

Godil, A., and Li, C., 2014 2014-06-12, en"Shrec'14 track: Extended large scale sketch-based 3d shape retrieval," (The Seventh Eurographics Workshop on 3D Object Retrieval (3DOR 2014)., Strasbourg, -1).

Hoffman, J., Tzeng, E., Park, T., Zhu, J.-Y., Isola, P., Saenko, K., Efros, A. A., and Darrell, T., 2017, "Cycada: Cycle-consistent adversarial domain adaptation,"

Jia, C., Yang, Y., Xia, Y., Chen, Y.-T., Parekh, Z., Pham, H., Le, Q. V., Sung, Y., Li, Z., and Duerig, T., 2021, "Scaling up visual and vision-language representation learning with noisy text supervision,"

Jiang, Z., Xu, F. F., Araki, J., and Neubig, G., 2020, "How can we know what language models know?." *Transactions of the Association for Computational Linguistics* **8**, 423–438.

Kaya, M., and Bilge, H. Ş., 2019, "Deep metric learning: A survey," *Symmetry* **11**, 1066.

Kishida, K., 2005, *Property of average precision and its generalization: An examination of evaluation indicator for information retrieval experiments* (National Institute of Informatics Tokyo, Japan).

Li, J., Chen, E., Ding, Z., Zhu, L., Lu, K., and Shen, H. T., 2021, "Maximum density divergence for domain adaptation," *IEEE Transactions on Pattern Analysis and Machine Intelligence* **43**, 3918–3930.

Long, M., Cao, Z., Wang, J., and Jordan, M. I., 2018, "Conditional adversarial domain adaptation," *Advances in neural information processing systems* **31**

Maturana, D., and Scherer, S., 2015, "Voxnet: A 3d convolutional neural network for real-time object recognition," in *2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 922–928.

Matusita, K., 1955, "Decision rules, based on the distance, for problems of fit, two samples, and estimation," *The Annals of Mathematical Statistics*, 631–640.

Peng, X., Bai, Q., Xia, X., Huang, Z., Saenko, K., and Wang, B., 2019, "Moment matching for multi-source domain adaptation," in *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 1406–1415.

Peng, X., Usman, B., Kaushik, N., Hoffman, J., Wang, D., and Saenko, K., 2017, "Visda: The visual domain adaptation challenge," *arXiv preprint arXiv:1710.06924*

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., *et al.*, 2021, "Learning transferable visual models from natural language supervision," in *International Conference on Machine Learning* (PMLR). pp. 8748–8763.

Saito, K., Watanabe, K., Ushiku, Y., and Harada, T., 2018, "Maximum classifier discrepancy for unsupervised domain adaptation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3723–3732.

Shin, T., Razeghi, Y., Logan IV, R. L., Wallace, E., and Singh, S., 2020, "Autoprompt: Eliciting knowledge from language models with automatically generated prompts," *arXiv preprint arXiv:2010.15980*

Sun, X., Wu, J., Zhang, X., Zhang, Z., Zhang, C., Xue, T., Tenenbaum, J. B., and Freeman, W. T., 2018, "Pix3d: Dataset and methods for single-image 3D shape modeling," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*

Venkateswara, H., Eusebio, J., Chakraborty, S., and Panchanathan, S., 2017, "Deep hashing network for unsupervised domain adaptation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 5018–5027.

Wang, F., Kang, L., and Li, Y., 2015, "Sketch-based 3d shape retrieval using convolutional neural networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1875–1883.

Wu, Z., Zhang, Y., Zeng, M., Qin, F., and Wang, Y., 2018, "Joint analysis of shapes and images via deep domain adaptation," *Computers Graphics* **70**, 140–147.

Xu, T., Chen, W., Wang, P., Wang, F., Li, H., and Jin, R., 2021, "Cdtrans: Cross-domain transformer for unsupervised domain adaptation," *arXiv preprint arXiv:2109.06165*

Yang, Y., Han, J., Zhang, D., and Cheng, D., 2021, "Disentangling deep network for reconstructing 3d object shapes from single 2d images," in *Pattern Recognition and Computer Vision*, edited by Ma, H., Wang, L., Zhang, C., Wu, F., Tan, T., Wang, Y., Lai, J., and Zhao, Y. (Springer International Publishing, Cham). pp. 153–166.

Yang, Y., Han, J., Zhang, D., and Tian, Q., 2022, "Exploring rich intermediate representations for reconstructing 3d shapes from 2d images," *Pattern Recognition* **122**, 108295.

Yao, T., Pan, Y., Ngo, C.-W., Li, H., and Mei, T., 2015, "Semi-supervised domain adaptation with subspace learning for visual recognition," in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2142–2150.

Yu, J., Yang, X., Gao, F., and Tao, D., 2016, "Deep multimodal distance metric learning using click constraints for image ranking," *IEEE transactions on cybernetics* **47**, 4014–4024.

Zhang, Y., Jiang, H., Miura, Y., Manning, C. D., and Langlotz, C. P., 2022, "Contrastive learning of medical visual representations from paired images and text," in *Machine Learning for Healthcare Conference* (PMLR). pp. 2–25.

Zhong, Z., Friedman, D., and Chen, D., 2021, "Factual probing is [mask]: Learning vs. learning to recall," *arXiv preprint arXiv:2104.05240*

Zhou, K., Liu, Z., Qiao, Y., Xiang, T., and Loy, C. C., 2022, "Domain generalization: A survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1–20.