DA-Ada: Learning Domain-Aware Adapter for Domain Adaptive Object Detection

Haochen Li[1,4] **Rui Zhang**[2*] **Hantao Yao**[3] **Xin Zhang**[2] **Yifan Hao**[2]
Xinkai Song[2] **Xiaqing Li**[2] **Yongwei Zhao**[2] **Ling Li**[1,4*] **Yunji Chen**[2,4]

[1]Intelligent Software Research Center, Institute of Software, CAS, Beijing, China
[2]State Key Lab of Processors, Institute of Computing Technology, CAS, Beijing, China
[3] State Key Laboratory of Multimodal Artificial Intelligence Systems, Institute of Automation, CAS, Beijing, China
[4] University of Chinese Academy of Sciences, Beijing, China
haochen2021@iscas.ac.cn, zhangrui@ict.ac.cn, hantao.yao@nlpr.ia.ac.cn, {zhangxin, haoyifan, songxinkai, lixiaqing, zhaoyongwei}@ict.ac.cn, liling@iscas.ac.cn, cyj@ict.ac.cn

Abstract

Domain adaptive object detection (DAOD) aims to generalize detectors trained on an annotated source domain to an unlabelled target domain. As the visual-language models (VLMs) can provide essential general knowledge on unseen images, freezing the visual encoder and inserting a domain-agnostic adapter can learn domain-invariant knowledge for DAOD. However, the domain-agnostic adapter is inevitably biased to the source domain. It discards some beneficial knowledge discriminative on the unlabelled domain, *i.e.*domain-specific knowledge of the target domain. To solve the issue, we propose a novel Domain-Aware Adapter (DA-Ada) tailored for the DAOD task. The key point is exploiting domain-specific knowledge between the essential general knowledge and domain-invariant knowledge. DA-Ada consists of the Domain-Invariant Adapter (DIA) for learning domain-invariant knowledge and the Domain-Specific Adapter (DSA) for injecting the domain-specific knowledge from the information discarded by the visual encoder. Comprehensive experiments over multiple DAOD tasks show that DA-Ada can efficiently infer a domain-aware visual encoder for boosting domain adaptive object detection. Our code is available at https://github.com/Therock90421/DA-Ada.

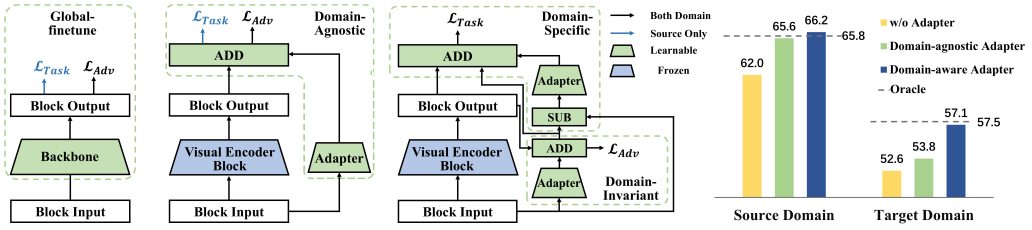
1 Introduction

Object detection [52, 51, 41, 12] have achieved remarkable performance, but suffer severe performance drop when dealing with unseen data due to domain discrepancy. To alleviate this problem, domain adaptive object detection (DAOD) [7] is explored to transfer a detector trained on the labelled source domain to a unlabelled target domain. Traditional DAOD works [7, 68, 82, 65, 78, 37, 71, 25] generate the domain-aligned feature via fine-tuning the backbone, as depicted in Fig. 1(a). Nevertheless, it is easily biased towards the source domain since a considerable number of parameters need to be updated with the annotations only from the source domain.

Recently, applying prompt tuning [81, 80, 73] on visual-language models (VLMs) is widely used for two reasons: 1) few parameters need to be learned; 2) VLMs trained on large-scale image-text pairs can extract highly generalized features. Recent works [31, 57] have explored using prompt tuning to generate the domain-aware detection head for DAOD. However, they all extract the visual feature

*Corresponding author.

38th Conference on Neural Information Processing Systems (NeurIPS 2024).



(a) Traditional Methods (b) Domain-agnostic Adapter (c) Domain-aware Adapter (d) mAP(%) Comparison

Figure 1: (a) Traditional DAOD methods optimize the backbone adversarially. (b) Domain-agnostic adapter is inserted into the frozen visual encoder to learn domain-invariant knowledge. (c) Domain-aware adapter can simultaneously capture the domain-specific knowledge from the discarded feature. (d) The mAP(%) comparison on the Cross-Weather Adaptation. Compared with original VLM, domain-agnostic adapter brings significant improvement to the source domain but limited improvement to the source domain, while domain-aware adapter brings significant improvement to both source domain and target domain.

from the image with a frozen visual encoder, ignoring learning task-related knowledge and limiting the improvement of visual features' discriminative capabilities.

To inject the task-related knowledge into the visual encoder, some methods [24, 8, 5] insert an adapter module into the frozen backbone. Formally, a adapter shared across domains can be straightly introduced to learn the task-related knowledge with the annotations of source domain and domain-aligned constraint, as shown in Fig. 1(b). However, this adapter is domain-agnostic and can only learn domain-invariant knowledge between the two domains under the domain-aligned constraint. Besides, the domain-invariant knowledge is inevitably biased to the source domain, since the annotations are only from the source domain. As shown in Fig. 1(d), compared with the original VLM, the domain-agnostic adapter brings significant improvement to the source domain, while the improvement of the target domain is limited. Summarily, the bias of the domain-invariant knowledge learned from domain-agnostic adapter limits the generalization to the unseen target domain.

Trained on large-scale data, the VLM provides essential general knowledge on unseen images, while the learned domain-invariant knowledge biased to the source domain shows limited improvement on the target domain. Consequently, when transferring essential general knowledge to the domain-invariant knowledge, the domain-agnostic adapter discards some beneficial knowledge on the target domain. Basically, it discards the domain-specific knowledge that distinguishes the target domain but is different from the domain-invariant knowledge. In summary, adding a complementary adapter to capture the target-specific knowledge from the discarded knowledge between the essential general knowledge and domain-invariant knowledge is an effective way to boost the performance of the VLM in DAOD task.

In this paper, we propose a novel Domain-aware Adapter (DA-Ada) to facilitate the visual encoder learning the domain-specific knowledge along with the domain-invariant knowledge. Formally, DA-Ada introduces a Domain-Invariant Adapter (DIA) and Domain-Specific Adapter (DSA) to exploit domain-invariant and domain-specific knowledge, respectively, as shown in Fig. 1(c). The DIA is attached to the block of the visual encoder in parallel and optimized by aligning the feature distribution of two domains to learn domain-invariant knowledge. The DSA is fed with the difference between the input and output of the block to recover the domain-specific knowledge discarded by the DIA. Since the difference represents the feature discarded by the block, the discarded knowledge between the essential general knowledge and domain-invariant knowledge is also hidden in the difference. Hence, the DSA can regain the domain-specific knowledge from the difference adaptively to improve the generalization ability on target domain, as shown in Fig. 1(d). Moreover, we propose the Visual-guided Textual Adapter (VTA), embedding cross-domain information learned by DA-Ada into textual encoder to enhance the discriminability of detection head. Overall, the proposed DA-Ada can inject domain-invariant and domain-specific knowledge into VLM for DAOD.

We conduct evaluations on mainstream DAOD benchmarks: Cross-Weather (Cityscapes $\rightarrow$ Foggy Cityscapes), Cross-Fov (KITTI $\rightarrow$ Cityscapes), Sim-to-Real (SIM10K $\rightarrow$ Cityscapes) and Cross-Style (Pascal VOC $\rightarrow$ Clipart). Experimental results show that the proposed DA-Ada brings noticeable improvement and outperforms state-of-the-art methods by a large margin. For example, DA-Ada reaches 58.5% mAP on Cross-Weather, surpassing the state-of-the-art DA-Pro [31] by 2.7%.

2 Related Work

Visual-Languague models Visual-language models (VLMs) [50, 33, 34] embed visual and text modalities into a shared space, enabling cross-modal alignment. Pre-trained with an astonishing scale of image-text pairs, they demonstrate comprehensive visual understanding. CLIP [50] simultaneously trains a visual encoder and a textual encoder with 400 million image-text pairs, showing promising performance on both the seen and unseen classes. Furthermore, [79, 19, 14, 61] distill the knowledge from the visual encoder of CLIP into the detection backbone and transform the textual encoder into detection head. Considering strong generalization, we apply RegionCLIP [79] as the detector.

Domain Adaptive Object Detection (DAOD) aims to adapt the object detector [52] trained on the labelled source domain to the unlabelled target domain. Previous approaches can be broadly divided into two orthogonal categories: feature alignment and semi-supervised learning. Feature alignment [45, 76, 46, 72, 56, 15, 47, 30, 76] aims to align the feature distributions of the two domains with domain discriminators [7], to generate domain-invariant knowledge in three levels: image-level [7, 68, 65, 40], instance-level [77, 53] and category-level [62, 25, 38]. To prevent knowledge unique to each domain from interfering with alignment, recent works [55, 2, 1, 32, 66] propose multiple extractors[42, 40, 63, 67, 27] and discriminators [69] to decouple the domain-invariant and domain-specific knowledge. In parallel, semi-supervised learning strives to augment training data with style transfer [43, 23, 10] and pseudo label [59, 39, 6, 11]. However, applying existing DAOD method to VLM would overfit the model to the training data, compromising the generalization of pre-trained models. To preserve the pre-trained knowledge, we opt to freeze the VLM and devise a novel domain-aware adapter to facilitate cross-domain adaptation. Compared with existing decoupling methods that only use domain-invariant features for detection, our method adopts a decoupling-refusion strategy. It adaptively modify domain-invariant features with domain-specific features to enhance the discriminability on the target domain.

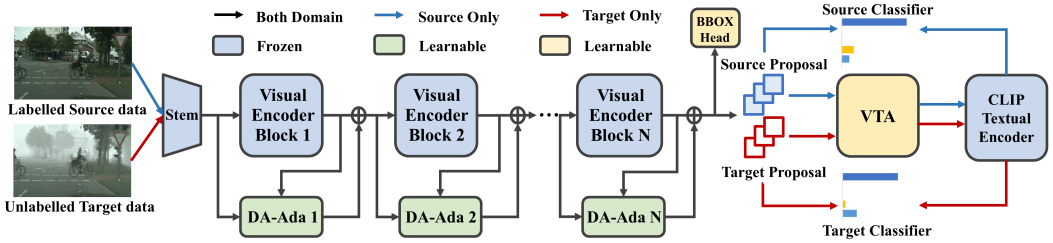
Tuning method for VLM Adapting pre-trained VLM to downstream tasks via global finetuning is prohibitively expensive and easily overfitted to training datasets. To solve this issue, prompt tuning [81] replaces the hand-crafted prompts with the learnable tokens for the textual encoder. Conditions like categories [80, 74], human prior [73] and domain knowledge [31] are attached to attain robust performance on new tasks. However, they freeze the visual encoder, preventing it from learning cross-domain information for DAOD.In parallel, originated from Natural Language Processing (NLP)[24, 49, 64, 85, 21], adapter tuning inserts learnable small layers into the visual encoder so that the backbone can learn knowledge from new tasks. ViT-Adapter [8] and Conv-Adapter [5] are proposed to efficiently transfer pre-trained knowledge to zero or few-shot visual tasks. [17] integrates the adapter into the CLIP model, and [58] further analyzes the components to be frozen or learnable. [48] combines self-supervised learning to enhance the ability to extract low-level features. Recent [70] explore injecting task-related knowledge into segmentation model SAM [29]. However, tuning the adapter directly on both domains will bias it towards the source domain and fails to distinguish domain-specific knowledge, leading to insufficient discrimination on the target domain. In this paper, we propose a novel domain-aware adapter that explicitly learns both domain-invariant and domain-specific knowledge to inject cross-domain information into the visual encoder.

3 Methodology

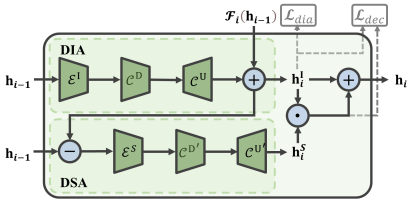
In this section, we present a novel Domain-aware Adapter (DA-Ada) tailored for DAOD. DA-Ada employs adapter tuning to introduce both domain-specific and domain-invariant knowledge into VLM. It is worth noting that the proposed method can be attached to any CNN-based detectors as a plug-and-play module. Without loss of generality, we take vanilla Faster-RCNN [52] as an example.

3.1 Overview

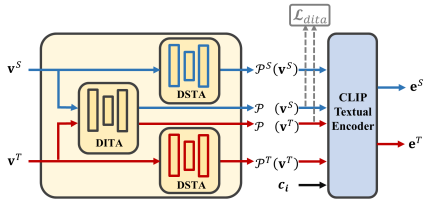
Inspired by adapter tuning, we can custom learnable adapters to inject cross-domain information into the visual encoder. Specifically, to enrich the extracted features with high domain generalization capabilities, an ideal adapter should satisfy conditions from the following two aspects. First, it can model the commonalities between the source and target domains, *i.e.*domain-invariant knowledge. Second, it can adaptively supply the unique attributes of each domain, *i.e.*domain-specific knowledge.



(a) Overview of DA-Ada



(b) Domain-aware Adapter Module (DA-Ada)



(c) Visual-guided Textual Adapter (VTA)

Figure 2: Overview of the proposed (a) DA-Ada for DAOD and the architecture of (b) the i-th domain-aware adapter module (c) the visual-guided textual adapter.

In this perspective, we design an effective Domain-Aware Adapter (DA-Ada) consisting of a Domain-Invariant Adapter (DIA) and a Domain-Specific Adapter (DSA). As shown in Fig. 2(a), given input image $\mathbf{x}$, we split the visual encoder into N blocks $\{\mathcal{F}_i\}_{i=1}^{N}$ by feature resolutions ($N = 4$ in ResNet). Then we attach N blocks with DA-Ada modules $\{\mathcal{A}_i\}_{i=1}^{N}$ in Fig. 2(b):

$$\mathbf{h}_0 = \mathcal{S}(\mathbf{x}); \mathbf{h}_i = \mathcal{A}_i(\mathbf{h}_{i-1}, \mathcal{F}_i(\mathbf{h}_{i-1})), \tag{1}$$

where $\mathcal{S}$ denotes the stem layer. For the i-th DA-Ada module, we first feed the i-th block's input $\mathbf{h}_{i-1}$ into the i-th DIA module $\mathcal{A}_i^I$ to extract the domain-invariant features $\mathbf{h}_i^I$. Then we attain the domain-specific features $\mathbf{h}_i^S$ from the subtraction of $\mathbf{h}_{i-1}$ and $\mathbf{h}_i^I$ by the DSA module $\mathcal{A}_i^S$:

$$\mathbf{h}_i^I = \mathcal{A}_i^I(\mathbf{h}_{i-1}) + \mathcal{F}_i(\mathbf{h}_{i-1}); \mathbf{h}_i^S = \mathcal{A}_i^S(\mathbf{h}_{i-1} - \mathbf{h}_i^I). \tag{2}$$

After that, we fuse $\mathbf{h}_i^I, \mathbf{h}_i^S$ with spatial attention to output $\mathbf{h}_i$ for i-th block:

$$\mathbf{h}_i = \mathbf{h}_i^I + \mathbf{h}_i^I \cdot \mathbf{h}_i^S, \tag{3}$$

where $\cdot$ denotes the element-wise Hadamard product. With N learnable adapters, we obtain visual embedding $\mathbf{v} = \mathbf{h}_N$ for subsequent detection. As the visual embedding contains sufficient cross-domain information, we propose the Visual-guided Textual Adapter (VTA), projecting the visual embedding to the textual encoder to enhance the discriminability of the detection head. As shown in Fig. 2(c), the visual-guided textual adapter uses the visual embedding $\mathbf{v}^S, \mathbf{v}^T$ to infer textual embedding $\mathbf{e}^S, \mathbf{e}^T$ on source and target domain, which is utilized for prediction. Overall, the proposed DA-Ada can inject domain-invariant and domain-specific knowledge into VLM to improve cross-domain generalization ability.

3.2 Domain-Invariant Adapter (DIA)

The DIA module is proposed to inject the domain-invariant knowledge into the visual encoder. As shown in Fig. 2(b), it applies a bottleneck to learn multi-scale domain knowledge, and the output distribution is aligned between domains for extracting domain-invariant knowledge. Specifically, for the i-th block of the visual encoder, we first forward the input feature $\mathbf{h}_{i-1} \in \mathbb{R}^{b\times c\times h\times w}$ to the i-th DIA and filter domain-irrelevant information with embedding block $\mathcal{E}^I$:

$$\mathbf{h}_i^E = \mathcal{E}^I(\mathbf{h}_{i-1}). \tag{4}$$

After that, the embedding $\mathbf{h}_i^E \in \mathbb{R}^{b\times c\times h\times w}$ is helpful for domain representation learning. Low channel-dimensional features have less information redundancy and are more suitable for domain adaptation than high-dimensional ones. Following this spirit, the embedding is encouraged to be down-projected to a low channel-dimensional vector $\mathbf{h}_i^L \in \mathbb{R}^{b\times r\times h\times w}$ to extract domain-invariant

knowledge and filter redundant information. Formally, a down-projection $\mathcal{C}^D$ is applied to reduce the dimension to r:

$$\mathbf{h}_i^L = \mathcal{C}^D(\mathbf{h}_i^E). \tag{5}$$

Considering that the scale of objects varies between domains, we introduce M down-projectors $\{\mathcal{C}_i^D\}_{i=1}^M$ with different receptive fields, enabling it to capture various spatial features across multiple scales. Specifically, the embedding $\mathbf{h}_i^E = [\mathbf{h}_{i,1}^E, \mathbf{h}_{i,2}^E, ..., \mathbf{h}_{i,M}^E]$ is first split up evenly in the channel dimension. Then, each partition is resized to different resolutions and down-projected. Therefore, the multi-scale version of Eq. (5) is expressed as:

$$\mathbf{h}_i^L = [\mathcal{C}_1^D(\mathbf{h}_{i,1}^E), \mathcal{C}_2^D(\mathbf{h}_{i,2}^E)..., \mathcal{C}_M^D(\mathbf{h}_{i,M}^E)]. \tag{6}$$

Furthermore, the low-dimensional knowledge $\mathbf{h}_i^L$ is encouraged to be mapped back to the original dimensional feature space and supplemented to the pre-trained features. Typically, we apply the dimension-raising function $\mathcal{C}^U$ on $\mathbf{h}_i^L$ to extract domain-invariant knowledge for the visual encoder.

$$\mathcal{A}_i^I(\mathbf{h}_{i-1}) = \mathcal{C}^U(\mathbf{h}_i^L), \tag{7}$$

where $\mathcal{A}_i^I(\mathbf{h}_{i-1}) \in \mathbb{R}^{b\times c\times h\times w}$ is output of the i-th DIA, and will be summed with $\mathcal{F}_i^I(\mathbf{h}_{i-1})$ to attain domain-invariant feature $\mathbf{h}_i^I$ in Eq. (2). To ensure DIA learning domain-invariant knowledge, the $\mathbf{h}_i^I$ is expected to be well aligned between the two domains. Therefore, N domain discriminator $\{\mathcal{D}_i\}_{i=1}^N$ is attached to each $\mathbf{h}_i^I$ to calculate adversarial loss $\mathcal{L}_{dia}$. We will introduce this loss in Sec.3.5.

With the combination of dimensional reduction-increase processes and the constraints of detection and adversarial loss, the DIA can extract domain-invariant features while reducing redundant features.

3.3 Domain-Specific Adapter (DSA)

Adapted with DIA, the essential general knowledge of the frozen VLM is transferred to domain-invariant knowledge. However, the knowledge learned only through the DIA is biased towards the source domain and appears less discriminative on the target domain. Considering the high generalization of essential general knowledge of the frozen VLM, we attribute this problem to the fact that the DIA discards some domain-specific knowledge that is highly generalizable on the unlabelled target domain. Since the difference between the input and output of the block denotes the discarded feature, the discarded domain-specific knowledge is also hidden in the difference. To this end, we propose the DSA module to recover domain-specific knowledge from the difference.

After the DIA injects the domain-invariant knowledge into the visual encoder, the domain-specific knowledge unique to the target domain is discarded by the output $\mathbf{h}_i^I$. Therefore, we first obtain the feature $\mathbf{h}_i^D$ discarded by the visual encoder block from the difference of the input $\mathbf{h}_{i-1}$ and $\mathbf{h}_i^I$:

$$\mathbf{h}_i^D = \mathcal{E}^S(\mathbf{h}_{i-1} - \mathbf{h}_i^I), \tag{8}$$

where $\mathcal{E}^S$ is similar with the embedding block $\mathcal{E}^I$. As domain-specific knowledge is hidden in the discarded difference $\mathbf{h}_i^D$, a bottleneck architecture is employed for adaptive knowledge extraction:

$$\mathbf{h}^{L'}{}_i = \mathcal{C}^{D'}(\mathbf{h}_i^D), \tag{9}$$

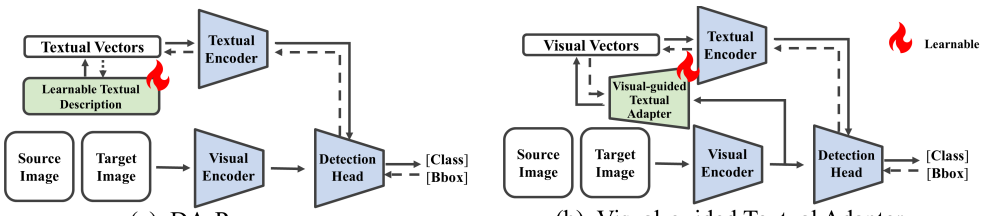
$$\mathbf{h}_i^S = \mathcal{A}_i^S(\mathbf{h}_{i-1} - \mathbf{h}_i^I) = \mathcal{C}^{U'}(\mathbf{h}^{L'}{}_i), \tag{10}$$

where $\mathcal{C}^{D'}, \mathcal{C}^{U'}$ follow the same configurations as $\mathcal{C}^D, \mathcal{C}^U$ to perceive multi-scale domain-specific knowledge in bottleneck manner.

Generally speaking, domain-invariant knowledge dominates the process of transferring essential general knowledge of the VLM, while domain-specific knowledge fine-tunes this process based on the characteristics of each domain. To this end, it is a more reasonable way to adaptively supplement domain-specific knowledge with the extracted $\mathbf{h}_i^S$ through pixel-level attention rather than straightforward addition. Therefore, the injection of the whole DA-Ada is written as Eq. (3).

3.4 Visual-guided Textual Adapter (VTA)

With the domain-aware adapter to inject domain-invariant and domain-specific knowledge, the extracted visual feature shows rich discriminability, which can also be used to improve detection



(a) DA-Pro (b) Visual-guided Textual Adapter

Figure 3: Comparison between (a) DA-Pro and (b) Visual-guided textual adapter.

head. Therefore, we introduce the VTA to exploit the cross-domain information contained in the visual features to enhance the textual encoder.

In order to fully exploit the domain-invariant and domain-specific knowledge extracted by the DA-Ada module, we equip two learnable components for VTA: the domain-invariant textual adapter $\mathcal{P}$ (DITA) and the domain-specific textual adapter $\mathcal{P}^S, \mathcal{P}^T$ (DSTA) shown in Fig. 2(c). The DITA is shared across domains to encode visual domain-invariant knowledge into the input of the textual encoder, optimized by a domain discriminator $\mathcal{D}^P$. The DSTA is tailored to further supplement domain-specific knowledge for the source domain S and the target domain T. In practice, the structure of DITA and DSTA is a 3-layer MLP with a hidden dimension of 512, projecting visual embeddings into 8 tokens for the textual encoder. Formally, the VTA embeds visual information into the textual embedding,

$$\mathbf{e}_i^S = \mathcal{T}(\mathcal{P}(\mathbf{v}^S), \mathcal{P}^S(\mathbf{v}^S), c_i); \mathbf{e}_i^T = \mathcal{T}(\mathcal{P}(\mathbf{v}^T), \mathcal{P}^T(\mathbf{v}^T), c_i), \tag{11}$$

where $\mathbf{v}^S, \mathbf{v}^T$, i and c_i denote the visual embedding from domain S, T, the i-th class and its textual description. $\mathcal{T}$ denotes the textual encoder. $\mathbf{e}_i^S$ and $\mathbf{e}_i^T$ is the textual-level classifier embedding of i-th class from the source and target domains, respectively.

Our proposed VTA introduces discriminative visual features into the textual encoder, alleviating the problem of insufficient adaptation in plain textual tuning. Existing methods [31] only tune learnable textual descriptions for detection head, as shown in Fig. 3(a). However, textual descriptions are insufficient to describe certain inter-domain differences, *e.g.*, differences in fields of view, leading to a limited ability to learn cross-domain information. Different from them, VTA analyses domain-invariant and domain-specific knowledge from visual features, inferring an image-conditional detection head with high discriminability, as shown in Fig. 3(b).

3.5 Optimization Objective

We aim to insert the DA-Ada into the visual encoder to learn cross-domain information and further tune the prompt for discriminative textual representation with image conditions. On the one hand, we introduce domain adversarial loss to the DIA and DITA to guide the learning of domain-invariant information. Formally, we obtain the output features $\mathbf{h}_i^{I,S}, \mathbf{h}_i^{I,T}$ for the source image $\mathbf{x}_s$ and the target image $\mathbf{x}_t$ of each DIA, and minimize the adversarial loss:

$$\mathcal{L}_{dia} = -\sum_{i=1}^{N}[\mathbb{E}_{\mathbf{x}_s}||\mathcal{D}_i(\mathbf{h}_i^{I,S})||_2^2 + \mathbb{E}_{\mathbf{x}_t}||\mathcal{D}_i(\mathbf{h}_i^{I,T}) - \mathbf{1}||_2^2]. \tag{12}$$

And the domain-shared DITA is expected to be aligned between domains:

$$\mathcal{L}_{dita} = -[\mathbb{E}_{\mathbf{x}_s}||\mathcal{D}^P(\mathbf{v}^S)||_2^2 + \mathbb{E}_{\mathbf{x}_t}||\mathcal{D}^P(\mathbf{v}^T) - \mathbf{1}||_2^2], \tag{13}$$

where $\mathbf{v}^S, \mathbf{v}^T$ denotes the source and target visual embedding.

On the other hand, we learn task-related domain-specific knowledge in a semi-supervised manner. For the source image, we calculate the cross-entropy for each visual embedding $\mathbf{v}^S$ with its annotations y. For the target $\mathbf{v}^T$, we first obtain the prediction via the hand-crafted prompt "A photo of [class]" and filter out high-confidence pseudo labels y', then minimize the cross-entropy as well:

$$\mathcal{L}_{det} = \mathcal{L}_{ce}(\mathbf{v}^S \times \mathbf{e}^S, y) + \mathcal{L}_{ce}(\mathbf{v}^T \times \mathbf{e}^T, y'), \tag{14}$$

where $\times$ denotes Matrix multiplication.

Table 1: Comparison (%) with existing methods on Cross-Weather Cityscapes→Foggy Cityscapes (C→F), Cross-Fov KITTI→Cityscapes (K→C) and Sim-to-Real adaptation SIM10K→Cityscapes (S→C). * denotes CLIP [50]-based methods.

	C→F									K→C	S→ C
Methods	Person	Rider	Car	Truck	Bus	Train	Motor	Bicycle	mAP	mAP	mAP
DA-Faster [7]	29.2	40.4	43.4	19.7	38.3	28.5	23.7	32.7	32.0	41.9	38.2
SIGMA++ [38]	46.4	45.1	61.0	32.1	52.2	44.6	34.8	39.9	44.5	49.5	57.7
CIGAR [44]	46.1	47.3	62.1	27.8	56.6	44.3	33.7	41.3	44.9	48.5	58.5
CSDA [16]	46.6	46.3	63.1	28.1	56.3	53.7	33.1	39.1	45.8	48.6	57.8
HT [11]	52.1	55.8	67.5	32.7	55.9	49.1	40.1	50.3	50.4	60.3	65.5
D^2-UDA [84]	46.9	53.3	64.5	38.9	61.0	48.5	42.6	54.2	50.6	60.3	58.1
AT [39]	56.3	51.9	64.2	38.5	45.5	55.1	**54.3**	35.0	50.9	-	-
NSA-UDA [83]	50.3	60.1	67.7	37.4	57.4	46.9	47.3	54.3	52.7	55.6	56.3
DA-Pro [31]*	55.4	62.9	70.9	40.3	63.4	54.0	42.3	58.0	55.9	61.4	62.9
DA-Ada(Ours)*	**57.8**	**65.1**	**71.3**	**43.1**	**64.0**	**58.6**	48.8	**58.7**	**58.5**	**66.7**	**67.3**

Table 2: Comparison (%) with existing methods on Cross-Style adaptation task Pascal VOC→Clipart. * denotes CLIP [50]-based methods.

Methods	Aero	Bike	Bird	Boat	Bottle	Bus	Car	Cat	Chair	Cow	Table	Dog	Horse	Motor	Person	Plant	Sheep	Sofa	Train	Tv	mAP
UaDAN [20]	35.0	73.7	41.0	24.4	21.3	69.8	53.5	2.3	34.2	61.2	31.0	**29.5**	47.9	63.6	62.2	61.3	13.9	7.6	48.6	23.9	40.2
FGRR [4]	30.8	52.1	35.1	32.4	42.2	62.8	42.6	21.4	42.8	58.6	33.5	20.8	37.2	81.4	66.2	50.3	21.5	29.3	**58.2**	47.0	43.3
UMT [10]	39.6	59.1	32.4	35.0	45.1	61.9	48.4	7.5	46.0	**67.6**	21.4	**29.5**	48.2	75.9	70.5	**56.7**	25.9	28.9	39.4	43.6	44.1
SIGMA [37]	40.1	55.4	37.4	31.1	54.9	54.3	46.6	23.0	44.7	65.6	23.0	22.0	42.8	55.6	67.2	55.2	32.9	**40.8**	45.0	58.6	44.5
TIA [78]	42.2	66.0	36.9	37.3	43.7	**71.8**	49.7	18.2	44.9	58.9	18.2	29.1	40.7	**87.8**	67.4	49.7	27.4	27.8	57.1	50.6	46.3
SIGMA++ [38]	36.3	54.6	40.1	31.6	58.0	60.4	46.2	**33.6**	44.4	66.2	25.7	25.3	44.4	58.8	64.8	55.4	36.2	38.6	54.1	**59.3**	46.7
CMT [60]	39.8	56.3	38.7	39.7	**60.4**	35.0	**56.0**	7.1	**60.1**	60.4	**35.8**	28.1	**67.8**	84.5	**80.1**	55.5	20.3	32.8	42.3	38.2	47.0
DA-Ada(Ours)*	**42.3**	**75.1**	**48.9**	**45.9**	49.0	**71.8**	55.6	15.4	50.7	56.6	19.9	20.6	61.3	80.7	73.0	29.2	**37.5**	21.5	52.5	52.9	**48.0**

Meanwhile, to decouple domain-invariant and domain-specific knowledge, we maximize the distribution discrepancy between DIA and DSA.

$$\mathcal{L}_{dec} = \sum_{i=1}^{N} [\mathbb{E}_{\mathbf{x}_s, \mathbf{x}_t} max[0, cos(\mathbf{h}_i^I, \mathbf{h}_i^I \cdot \mathbf{h}_i^S) - \beta]], \tag{15}$$

where $cos(\mathbf{a}, \mathbf{b}) = \frac{|\mathbf{a}^\top \cdot \mathbf{b}|}{||\mathbf{a}||_2^2 ||\mathbf{b}||_2^2}$ is absolute value of cosine distance, β is a threshold.

With the help of domain classifiers, DIA and DITA are encouraged to contain more domain-invariant knowledge. By minimizing $\mathcal{L}_{dec}$, the gap between DIA and DSA will be enlarged, which promotes DSA to extract more domain-specific knowledge. Overall, the optimization objective is:

$$\mathcal{L} = \mathcal{L}_{det} + \lambda_{dia}\mathcal{L}_{dia} + \lambda_{dita}\mathcal{L}_{dita} + \lambda_{dec}\mathcal{L}_{dec} + \mathcal{L}_{reg}, \tag{16}$$

where $\mathcal{L}_{reg}$ is the regression loss, and $\lambda_{dia}, \lambda_{dita}, \lambda_{dec}$ are balance ratios.

4 Experiment

4.1 Datasets and Implementation

We evaluate our method on four benchmarks: Cross-Weather(Cityscapes [9]→Foggy Cityscapes [54]), Cross-Fov(KITTI [18]→Cityscapes), Sim-to-Real(SIM10k [28]→Cityscapes) and Cross-Style(Pascal VOC [13]→Clipart [26]). Following [31], we adapt RegionCLIP(ResNet-50 [22]) with Faster-RCNN architecture as the baseline detector. We detail the datasets and implementation in Sec. 6.1and 6.3 of the Appendix.

4.2 Comparison to SOTA methods

We present representative state-of-the-art DAOD approaches for comparison, including feature alignment and semi-supervised learning methods.

Cross-Weather Adaptation Scenario Table 1 (C→F) illustrates that the proposed DA-Ada surpasses SOTA DA-Pro [31] by a remarkable margin of 2.6%, achieving the highest mAP over eight classes of 58.5%. Compared with existing methods, DA-Ada significantly improves seven categories (*i.e.* person,

Table 3: Comparison (%) of domain-aware adapter with standard adapter.

Adapter	Source-only	Domain-agnostic	Domain-aware
Cross-Weather	50.4	53.8	**57.1**
Cross-FoV	57.9	63.2	**65.8**
Sim-to-Real	58.4	62.4	**65.3**
Cross-Style	38.3	44.0	**46.4**

Table 4: Ablation studies (%) of domain-aware adapter on Cross-Weather adaptation.

DIA	$\mathcal{L}_{dia}$	DSA	$\mathcal{L}_{dec}$	mAP	Gains
				52.6	-
✓				53.8	+1.2
✓	✓			54.8	+2.2
✓	✓	✓		56.2	+3.6
✓	✓	✓	✓	**57.1**	**+4.5**

Table 5: Ablation (%) on insertion site of domain-aware adapter on Cross-Weather adaptation.

Block 1	Block 2	Block 3	Block 4	mAP
			✓	53.6
		✓		54.0
	✓			54.6
✓				55.1
✓	✓			56.9
✓	✓	✓		57.7
✓	✓	✓	✓	**58.5**

Table 6: Ablation (%) on input and injection operation of domain-aware adapter on Cross-Weather adaptation.

Input of DIA	Input of DSA	Injection Operation	mAP
		$\mathcal{F}_i(\mathbf{h}_{i-1})$	52.6
$\mathbf{h}_{i-1}$		$\mathbf{h}_i^I = \mathcal{F}_i(\mathbf{h}_{i-1}) + \mathcal{A}_i(\mathbf{h}_{i-1})$	54.8
$\mathbf{h}_{i-1}$	$\mathcal{F}_i(\mathbf{h}_{i-1})$	$\mathbf{h}_i^I + \mathbf{h}_i^S$	55.2
$\mathbf{h}_{i-1}$	$\mathbf{h}_{i-1}$ - $\mathcal{F}_i(\mathbf{h}_{i-1})$	$\mathbf{h}_i^I + \mathbf{h}_i^S$	56.2
$\mathbf{h}_{i-1}$	$\mathbf{h}_{i-1}$ - $\mathbf{h}_i^I$	$\mathbf{h}_i^I + \mathbf{h}_i^S$	56.7
$\mathbf{h}_{i-1}$	$\mathbf{h}_{i-1}$ - $\mathbf{h}_i^I$	Cross-Attention($\mathbf{h}_i^I$, $\mathbf{h}_i^S$, $\mathbf{h}_i^S$)	57.0
$\mathbf{h}_{i-1}$	$\mathbf{h}_{i-1}$ - $\mathbf{h}_i^I$	$\mathbf{h}_i^I + \mathbf{h}_i^I \cdot \mathbf{h}_i^S$	**57.1**

Table 7: Ablation (%) on Bottleneck dimension of domain-aware adapter. * denotes input dimension.

Bottleneck Dimension				mAP
DA-Ada 1	DA-Ada 2	DA-Ada 3	DA-Ada 4	
16	64	128	256	55.9
32	128	256	512	**57.1**
48	192	384	768	56.8
64*	256*	512*	1024*	56.6

Table 8: Comparison (%) of VTA with plain textual tuning methods.

Methods	C→F	Gains	K→C	Gains
Hand-crafted Prompt [79]	52.6	-	59.5	-
COOP [81]	53.5	+0.9	60.7	+1.2
DA-Pro [31]	55.1	+2.5	61.4	+1.9
VTA(Ours)	**55.8**	**+3.2**	**62.9**	**+3.4**

rider, car, truck, bus, train, and bicycle) ranging from 0.4% to 5.3%. The superior performance shows the remarkable effectiveness of the DA-Ada in the cross-domain generalization ability.

Cross-FOV Adaptation Scenario Table 1 (K→C) indicates a noticeable 5.3% improvement on the SOTA DA-Pro [31] by the DA-Ada. As K→C adaptation faces more complicated shape confusion than C→F, it requests higher discriminability of the model. Therefore, the considerable enhancement validates that the DA-Ada can efficiently learn robust visual encoder.

Sim-to-Real Adaptation Scenario We report the experimental results on SIM10k → Cityscapes benchmark in Table 1 (S→C). The proposed DA-Ada achieves the best results of 67.3% mAP, outperforming the previous best entry HT [11] 65.5% with 1.8%. The performance of DA-Ada is superior in the difficult adaptation task, which further demonstrates that our strategy is robust not only in appearance but also in more complex semantics adaptation tasks.

Cross-Style Adaptation Scenario Additionally, we assess DA-Ada on the more challenging Cross-Style adaptation, where the semantic hierarchy has a broader domain gap. DA-Ada peaks with 48.0%, outperforming all the SOTA methods presented in Table 2, demonstrating that injecting cross-domain information into the visual encoder could benefit the adaptation. Especially, DA-Ada exceeds all the compared methods on six categories (aeroplane, bike, bird, boat, bus, and sheep), which verifies the method is effective under challenging domain shifts and in multi-class problem scenarios.

4.3 Ablation Studies

Standard Adapter *vs.* Domain-aware Adapter We first compare the performance of the domain-aware adapter with existing adapters, including source-only adapter and domain-agnostic adapter. As shown in Table 3, while the domain-agnostic adapter surpasses the source-only version by $3.4\% \sim 5.7\%$ on four benchmarks, applying the domain-aware adapter further improves $2.4\% \sim 3.3\%$ mAP. We further explored the reasons for this advantage, shown in Fig 1(d). Compared with oracle, the domain-agnostic adapter reaches similar performance on the source domain, but suffers severe performance drop of 3.7% on the target domain, indicating that it is biased towards the source domain. While improving the source domain with 0.4%, our method reaches the oracle on the target domain. The superioir performance indicates the domain-aware adapter not only aligns domain-invariant knowledge more accurately, but also utilizes domain-specific knowledge to improve the detector's discriminative ability on the target domain.

Ablation for Domain-aware Adapter We conduct comprehensive ablation studies on each component of the proposed method in Table 4. Only introducing DIA to the backbone attains an mAP of 53.8%, and optimizing each DIA with an independent discriminator $\mathcal{L}_{dia}$ increases 1.0%. This indicates that learning domain-invariant adapters transfer task-related source knowledge to the target

Table 9: Ablation studies (%) of VTA on Cross-Weather adaptation.

DITA	$\mathcal{L}_{dita}$	DSTA	mAP	Gains
			57.1	-
✓			57.6	+0.5
✓	✓		57.9	+0.8
✓	✓	✓	**58.5**	**+1.4**

Table 10: Comparison (%) of computational efficiency on Cross-Weather adaptation

Method	Backbone Param (M)	Learnable Param (M)	mAP	Abs. Gains
DSS [65]	29.812	29.812	40.9	+4.2
CSDA [16]	33.645	33.645	45.3	+6.9
AT [39]	**39.225**	18.723	50.9	+7.9
DA-Pro [31]	34.834	**0.008**	55.9	+3.3
DA-Ada(Ours)	36.620	1.794	**58.5**	**+8.0**

domain. Moreover, the DSA boosts the DIA by 1.4% and 2.3% with the help of $\mathcal{L}_{dec}$, showing that learning domain-specific knowledge improves the discrimination of the target detection head.

Insertion Site We explicitly study the insertion site of DA-Ada, as shown in Table 5. When single adapter is applied, inserting the DA-Ada in the shallow block achieves better performance, *e.g.*DA-Ada with block 1 obtain 55.1%, surpassing all other insertion sites with block 2/3/4. And increasing the number of DA-Ada from 1 to 4 leads to steady improvements of $1.8\%, 0.8\%, 0.8\%$ respectively.

Input and Injection Operation We analyze different input features and injection operations of DIA/DSA in Table 6. Directly inserting DIA into the visual encoder and directly adding to the output of each block attains 2.2% improvement, showing the effectiveness of learning domain-invariant knowledge. However, there is limited performance gain in sending the output $\mathcal{F}_i(\mathbf{h}_{i-1})$ to DSA. It indicates that domain-specific knowledge is ignored during feature extraction of the visual encoder. To this end, inputting $\mathbf{h}_{i-1} - \mathcal{F}_i(\mathbf{h}_{i-1})$ to DSA receives 56.2%, exhibiting that the DSA can regain the domain-specific knowledge from the difference. As $\mathbf{h}_i^I = \mathcal{A}_i^I(\mathbf{h}_{i-1}) + \mathcal{F}_i(\mathbf{h}_{i-1})$ is updated to be domain-invariant, $\mathbf{h}_{i-1}$ - ($\mathbf{h}_i^I$) removes domain-invariant parts and appears to be domain-specific. Therefore, we forward it to DSA and gain an improvement of 0.5%, demonstrating the efficacy of learning domain-specific knowledge. Additionally, we substitute cross-attention and pixel-level attention for the direct addition, and gains highest mAP of 57.0% and 57.1%. It reveals that domain-specific knowledge describes intra-domain properties and is more suitable for refining the extracted features. For efficiency, we adopt the simpler pixel-level attention as the fusion function.

Bottleneck Dimension We also conduct an ablation study in Table 7 to explore the optimal bottleneck dimension of the DA-Ada. As the dimension increases, the performance peaks 57.1% when the bottleneck dimension is $1/2$ of the input and then appears to decline. We conclude that appropriate dimensional reduction can filter redundant features while extracting task knowledge.

Textual Tuning *vs.* Visual-guided Textual Adapter We compare the visual-guided textual adapter against existing methods, as shown in Table 8. Guided by visual conditions, VTA outperforms SOTA plain textual tuning methods by margins of 0.7% and 1.5% in two scenarios. Notably, VTA excels in the challenging Cross-FoV adaptation, suggesting that the visual modality effectively supplements the limitations of the textual encoder in describing domain attributes.

Ablation for Visual-guided Textual Adapter As shown in Table 9, learning DITA attains an mAP of 57.6%, and by introducing an additional adversarial loss, it achieves 57.9%. Moreover, with DSTA generating prompts for each domain, it exhibits a full adaptation performance of 58.5%. This shows that embedding image conditions into textual encoder can promote cross-domain detection.

Computational Efficiency As shown in Table 10, employing VLM yields a similar parameter scale while achieving a peak mAP of 58.5%. This indicates that the superior performance of DA-Ada does not arise from an increase in parameters. Furthermore, DA-Ada achieves the highest absolute gain of $+8.0\%$ with the training of only 1.794M parameters, demonstrating remarkable efficiency.

4.4 Detection Visualization

In Fig. 4, we present the comparison of the ground truth boxes (a) and the detection boxes of SOTA DA-Pro [31](c) and our methods (b)(d) on the target domain. (a.1)(b.1)(c.1)(d.1) are zoomed from the same region of images (a)(b)(c)(d) for a better view. Fig. 4(a.1) presents eight objects in the cropped region: 6 overlapped cars and a rider with a bicycle. The baseline model only detects two clear cars in Fig. 4(b.1). Failing to describe domain information, like weather conditions, it misses other objects hidden in the fog. In Fig. 4(c), the DA-Pro distinguishes the rider and the bicycle and improves 9.3% mAP with the domain-adaptive prompt. However, it ignores one car on the left of Fig. 4(c.1), suffering limited generalization ability due to insufficient domain representation learning in the visual encoder. Our proposed DA-Ada correctly detects the missing car (labelled in green) in the cropped region Fig. 4(d.1). By injecting cross-domain information into the visual encoder,

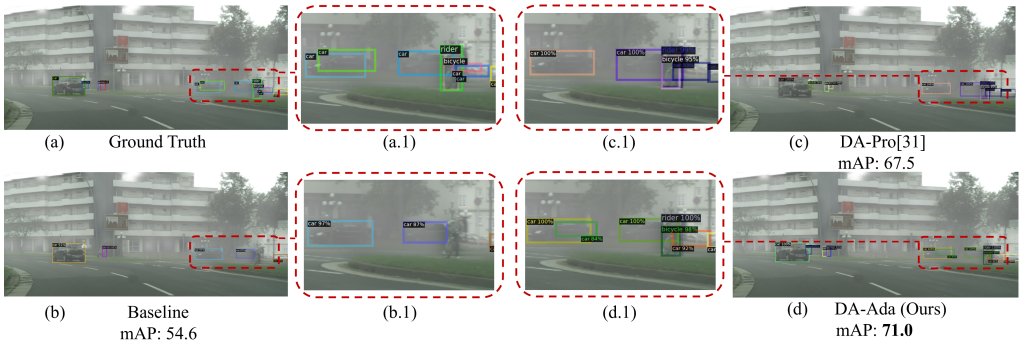


Figure 4: Detection comparison on the Cross-Weather adaptation scenario. We visualize the ground truth (a), the detection boxes of SOTA DA-Pro [31](c) and our methods (b)(d). mAP: mean Average Precision on the example image

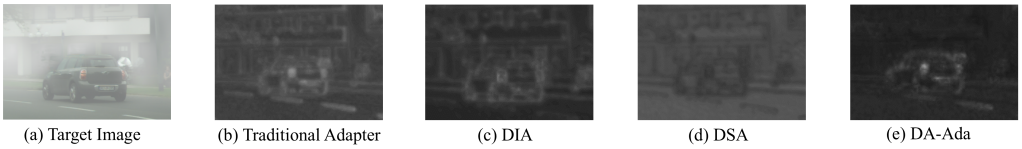


Figure 5: Feature comparison on the Cross-Weather adaptation scenario. We visualize (a) the target image and the output feature of (b) traditional adapter, (b) domain-invariant adapter (DIA), (c) domain-specific adapter (DSA) and (d) domain-aware adapter (DA-Ada).

the DA-Ada enables the model to detect more confidently on two bicycles $(83\%, 98\%)$ and one person (91%), compared with DA-Pro's $(79\%, 95\%)$ and (89%). These comparison results reveal the effectiveness of DA-Ada.

4.5 Feature Visualization

In Fig. 5, we visualize the output features of the traditional adapter, the domain-invariant adapter (DIA), domain-specific adapter (DSA) and the domain-aware adapter (DA-Ada). We sample image (a) a car and a person in the fog from Foggy Cityscapes. The traditional adapter (b) roughly extracts the outline of the car. However, affected by target domain attributes, such as fog, background areas are also highlighted in (b), and the person is not salient. DIA (c) mainly focuses on the object area and extracts domain-shared task information. DSA (d) mainly focuses on factors related to domain attributes besides the objects, such as foggy areas. By combining DIA with DSA, DA-Ada (e) extracts the car and person while reducing the interference of fog in the background. Compared with (b), objects are more salient in (e), indicating the effectiveness of DA-Ada.

5 Conclusion

In this paper, we propose a novel Domain-Aware Adapter (DA-Ada) for DAOD. As a small learnable attachment, it transfers highly generalized knowledge the visual-language model provides to cross-domain information for DAOD. Precisely, it consists of a Domain-Invariant Adapter (DIA) for learning domain-invariant knowledge and a Domain-Specific Adapter (DSA) for recovering the domain-specific knowledge from information discarded by the visual encoder. Extensive experiments over multiple DAOD tasks validate the effectiveness of DA-Ada in inferring a discriminative detector.

Acknowledgments and Disclosure of Funding

This work is partially supported by the National Key R&D Program of China(under Grant 2023YFB4502200), the NSF of China(under Grants 92364202, 61925208, U22A2028, 62222214, 62341411, 62102398, 62102399, U20A20227, 62372436, 62302478, 62302482, 62302483, 62302480, 62376268), 2022 Fundamental Disciplines Top Quality Student Training Program 2.0 Project, Strategic Priority Research Program of the Chinese Academy of Sciences, (Grant No.XDB0660200, XDB0660201, XDB0660202), Major Program of ISCAS(Grant No. ISCAS-ZD-202402), Beijing Natural Science Foundation(4222039), CAS Project for Young Scientists in Basic Research(YSBR-029), Youth Innovation Promotion Association CAS and Xplore Prize.

References

[1] Konstantinos Bousmalis, George Trigeorgis, Nathan Silberman, Dilip Krishnan, and Dumitru Erhan. Domain separation networks. In *NeurIPS*, pages 343–351, 2016.

[2] Woong-Gi Chang, Tackgeun You, Seonguk Seo, Suha Kwak, and Bohyung Han. Domain-specific batch normalization for unsupervised domain adaptation. In *CVPR*, pages 7354–7362, 2019.

[3] Chaoqi Chen, Jiongcheng Li, Zebiao Zheng, Yue Huang, Xinghao Ding, and Yizhou Yu. Dual bipartite graph learning: A general approach for domain adaptive object detection. In *ICCV*, pages 2703–2712, 2021.

[4] Chaoqi Chen, Jiongcheng Li, Hong-Yu Zhou, Xiaoguang Han, Yue Huang, Xinghao Ding, and Yizhou Yu. Relation matters: foreground-aware graph-based relational reasoning for domain adaptive object detection. *TPAMI*, 45(03):3677–3694, 2023.

[5] Hao Chen, Ran Tao, Han Zhang, Yidong Wang, Wei Ye, Jindong Wang, Guosheng Hu, and Marios Savvides. Conv-adapter: Exploring parameter efficient transfer learning for convnets. *arXiv preprint arXiv:2208.07463*, 2022.

[6] Meilin Chen, Weijie Chen, Shicai Yang, Jie Song, Xinchao Wang, Lei Zhang, Yunfeng Yan, Donglian Qi, Yueting Zhuang, Di Xie, et al. Learning domain adaptive object detection with probabilistic teacher. In *ICML*, pages 3040–3055, 2022.

[7] Yuhua Chen, Wen Li, Christos Sakaridis, Dengxin Dai, and Luc Van Gool. Domain adaptive faster r-cnn for object detection in the wild. In *CVPR*, pages 3339–3348, 2018.

[8] Zhe Chen, Yuchen Duan, Wenhai Wang, Junjun He, Tong Lu, Jifeng Dai, and Yu Qiao. Adapterfusion: Non-destructive task composition for transfer learning. *arXiv preprint arXiv:2205.08534*, 2022.

[9] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *CVPR*, 2016.

[10] Jinhong Deng, Wen Li, Yuhua Chen, and Lixin Duan. Unbiased mean teacher for cross-domain object detection. In *CVPR*, pages 4091–4101, 2021.

[11] Jinhong Deng, Dongli Xu, Wen Li, and Lixin Duan. Harmonious teacher for cross-domain object detection. In *CVPR*, pages 23829–23838, 2023.

[12] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2021.

[13] Mark Everingham, SM Ali Eslami, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes challenge: A retrospective. *IJCV*, 111:98–136, 2015.

[14] Mohammad Fahes, Tuan-Hung Vu, Andrei Bursuc, Patrick Pérez, and Raoul De Charette. Poda: Prompt-driven zero-shot domain adaptation. In *ICCV*, pages 18623–18633, 2023.

[15] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. Domain-adversarial training of neural networks. *JMLR*, 17(1): 2096–2030, 2016.

[16] Changlong Gao, Chengxu Liu, Yujie Dun, and Xueming Qian. Csda: Learning category-scale joint feature for domain adaptive object detection. In *ICCV*, pages 11421–11430, 2023.

[17] 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. *IJCV*, 132(2):581–595, 2024.

[18] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *CVPR*, 2012.

[19] Xiuye Gu, Tsung-Yi Lin, Weicheng Kuo, and Yin Cui. Open-vocabulary object detection via vision and language knowledge distillation. In *ICLR*, 2022.

[20] Dayan Guan, Jiaxing Huang, Aoran Xiao, Shijian Lu, and Yanpeng Cao. Uncertainty-aware unsupervised domain adaptation in object detection. *TMM*, 24:2502–2514, 2021.

[21] Junxian He, Chunting Zhou, Xuezhe Ma, Taylor Berg-Kirkpatrick, and Graham Neubig. Towards a unified view of parameter-efficient transfer learning. In *ICLR*, 2021.

[22] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016.

[23] Mengzhe He, Yali Wang, Jiaxi Wu, Yiru Wang, Hanqing Li, Bo Li, Weihao Gan, Wei Wu, and Yu Qiao. Cross domain object detection by target-perceived dual branch distillation. In *CVPR*, pages 9570–9580, 2022.

[24] Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. Parameter-efficient transfer learning for nlp. In *ICML*, pages 2790–2799, 2019.

[25] Jiaxing Huang, Dayan Guan, Aoran Xiao, Shijian Lu, and Ling Shao. Category contrast for unsupervised domain adaptation in visual tasks. In *CVPR*, pages 1203–1214, 2022.

[26] Naoto Inoue, Ryosuke Furuta, Toshihiko Yamasaki, and Kiyoharu Aizawa. Cross-domain weakly-supervised object detection through progressive domain adaptation. In *CVPR*, pages 5001–5009, 2018.

[27] Yifan Jiao, Hantao Yao, Bing-Kun Bao, and Changsheng Xu. Source-guided target feature reconstruction for cross-domain classification and detection. *TIP*, 2024.

[28] Matthew Johnson-Roberson, Charles Barto, Rounak Mehta, Sharath Nittur Sridhar, Karl Rosaen, and Ram Vasudevan. Driving in the matrix: Can virtual worlds replace human-generated annotations for real world tasks? In *ICRA 2017*, pages 746–753. IEEE, 2017.

[29] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *ICCV*, pages 4015–4026, 2023.

[30] Congcong Li, Dawei Du, Libo Zhang, Longyin Wen, Tiejian Luo, Yanjun Wu, and Pengfei Zhu. Spatial attention pyramid network for unsupervised domain adaptation. In *ECCV*, pages 481–497. Springer, 2020.

[31] Haochen Li, Rui Zhang, Hantao Yao, Xinkai Song, Yifan Hao, Yongwei Zhao, Ling Li, and Yunji Chen. Learning domain-aware detection head with prompt tuning. *NeurIPS*, 36, 2024.

[32] Haochen Li, Rui Zhang, Hantao Yao, Xin Zhang, Yifan Hao, Xinkai Song, and Ling Li. React: Remainder adaptive compensation for domain adaptive object detection. *TIP*, 33:3735–3748, 2024.

[33] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. BLIP: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *ICML*, pages 12888–12900, 2022.

[34] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*, 2023.

[35] Kai Li, Curtis Wigington, Chris Tensmeyer, Vlad I. Morariu, Handong Zhao, Varun Manjunatha, Nikolaos Barmpalios, and Yun Fu. Improving cross-domain detection with self-supervised learning. In *CVPR*, pages 4745–4754, 2023.

[36] Wuyang Li, Xinyu Liu, Xiwen Yao, and Yixuan Yuan. Scan: Cross domain object detection with semantic conditioned adaptation. In *AAAI*, volume 6, page 7, 2022.

[37] Wuyang Li, Xinyu Liu, and Yixuan Yuan. Sigma: Semantic-complete graph matching for domain adaptive object detection. In *CVPR*, pages 5291–5300, 2022.

[38] Wuyang Li, Xinyu Liu, and Yixuan Yuan. Sigma++: Improved semantic-complete graph matching for domain adaptive object detection. *TPAMI*, 45(07):9022–9040, 2023.

[39] Yu-Jhe Li, Xiaoliang Dai, Chih-Yao Ma, Yen-Cheng Liu, Kan Chen, Bichen Wu, Zijian He, Kris Kitani, and Peter Vajda. Cross-domain adaptive teacher for object detection. In *CVPR*, pages 7581–7590, 2022.

[40] Chuang Lin, Zehuan Yuan, Sicheng Zhao, Peize Sun, Changhu Wang, and Jianfei Cai. Domain-invariant disentangled network for generalizable object detection. In *ICCV*, pages 8771–8780, 2021.

[41] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *CVPR*, pages 2117–2125, 2017.

[42] Dongnan Liu, Chaoyi Zhang, Yang Song, Heng Huang, Chenyu Wang, Michael Barnett, and Weidong Cai. Decompose to adapt: Cross-domain object detection via feature disentanglement. *TMM*, 25:1333–1344, 2022.

[43] Rui Liu, Yahong Han, Yaowei Wang, and Qi Tian. Frequency spectrum augmentation consistency for domain adaptive object detection. *arXiv preprint arXiv:2112.08605*, 2021.

[44] Yabo Liu, Jinghua Wang, Chao Huang, Yaowei Wang, and Yong Xu. Cigar: Cross-modality graph reasoning for domain adaptive object detection. In *CVPR*, pages 23776–23786, 2023.

[45] Mingsheng Long, Yue Cao, Jianmin Wang, and Michael Jordan. Learning transferable features with deep adaptation networks. In *ICML*, pages 97–105. PMLR, 2015.

[46] Mingsheng Long, Han Zhu, Jianmin Wang, and Michael I Jordan. Unsupervised domain adaptation with residual transfer networks. *NeurIPS*, 29, 2016.

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

[48] Omiros Pantazis, Gabriel Brostow, Kate Jones, and Oisin Mac Aodha. Svl-adapter: Self-supervised adapter for vision-language pretrained models. *arXiv preprint arXiv:2210.03794*, 2022.

[49] Jonas Pfeiffer, Aishwarya Kamath, Andreas Rücklé, Kyunghyun Cho, and Iryna Gurevych. Adapterfusion: Non-destructive task composition for transfer learning. *arXiv preprint arXiv:2005.00247*, 2020.

[50] 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 *ICML*, pages 8748–8763. PMLR, 2021.

[51] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *CVPR*, pages 779–788, 2016.

[52] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *NeurIPS*, 28, 2015.

[53] Kuniaki Saito, Yoshitaka Ushiku, Tatsuya Harada, and Kate Saenko. Strong-weak distribution alignment for adaptive object detection. In *CVPR*, pages 6956–6965, 2019.

[54] Christos Sakaridis, Dengxin Dai, and Luc Van Gool. Semantic foggy scene understanding with synthetic data. *IJCV*, 126(9):973–992, Sep 2018.

[55] Sunandini Sanyal, Ashish Ramayee Asokan, Suvaansh Bhambri, Akshay Kulkarni, Jogendra Nath Kundu, and R Venkatesh Babu. Domain-specificity inducing transformers for source-free domain adaptation. In *ICCV*, pages 18928–18937, 2023.

[56] Rui Shao, Xiangyuan Lan, and Pong C Yuen. Feature constrained by pixel: Hierarchical adversarial deep domain adaptation. In *ACMMM*, pages 220–228, 2018.

[57] Mainak Singha, Harsh Pal, Ankit Jha, and Biplab Banerjee. Ad-clip: Adapting domains in prompt space using clip. In *ICCV*, pages 4355–4364, 2023.

[58] Yi-Lin Sung, Jaemin Cho, and Mohit Bansal. Vl-adapter: Parameter-efficient transfer learning for vision-and-language tasks. In *CVPR*, pages 5227–5237, 2022.

[59] Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. *NPIS*, 30, 2017.

[60] VS Vibashan, Poojan Oza, and Vishal M Patel. Instance relation graph guided source-free domain adaptive object detection. In *CVPR*, pages 3520–3530, 2023.

[61] Vidit Vidit, Martin Engilberge, and Mathieu Salzmann. Clip the gap: A single domain generalization approach for object detection. In *CVPR*, pages 3219–3229, 2023.

[62] Vibashan Vs, Vikram Gupta, Poojan Oza, Vishwanath A Sindagi, and Vishal M Patel. Mega-cda: Memory guided attention for category-aware unsupervised domain adaptive object detection. In *CVPR*, pages 4516–4526, 2021.

[63] Haoan Wang, Shilong Jia, Tieyong Zeng, Guixu Zhang, and Zhi Li. Triple feature disentanglement for one-stage adaptive object detection. In *AAAI*, pages 5401–5409, 2024.

[64] Ruize Wang, Duyu Tang, Nan Duan, Zhongyu Wei, Xuanjing Huang, Guihong Cao, Daxin Jiang, Ming Zhou, et al. K-adapter: Infusing knowledge into pre-trained models with adapters. *arXiv preprint arXiv:2002.01808*, 2020.

[65] Yu Wang, Rui Zhang, Shuo Zhang, Miao Li, Yangyang Xia, XiShan Zhang, and ShaoLi Liu. Domain-specific suppression for adaptive object detection. In *CVPR*, pages 9603–9612, 2021.

[66] Aming Wu and Cheng Deng. Single-domain generalized object detection in urban scene via cyclic-disentangled self-distillation. In *CVPR*, pages 847–856, 2022.

[67] Aming Wu, Yahong Han, Linchao Zhu, and Yi Yang. Instance-invariant domain adaptive object detection via progressive disentanglement. *TPAMI*, 44(8):4178–4193, 2021.

[68] Aming Wu, Rui Liu, Yahong Han, Linchao Zhu, and Yi Yang. Vector-decomposed disentanglement for domain-invariant object detection. In *ICCV*, pages 9342–9351, 2021.

[69] Jiaxi Wu, Jiaxin Chen, Mengzhe He, Yiru Wang, Bo Li, Bingqi Ma, Weihao Gan, Wei Wu, Yali Wang, and Di Huang. Target-relevant knowledge preservation for multi-source domain adaptive object detection. In *CVPR*, pages 5301–5310, 2022.

[70] Junde Wu, Wei Ji, Yuanpei Liu, Huazhu Fu, Min Xu, Yanwu Xu, and Yueming Jin. Medical sam adapter: Adapting segment anything model for medical image segmentation. *arXiv preprint arXiv:2304.12620*, 2023.

[71] Minghao Xu, Hang Wang, Bingbing Ni, Qi Tian, and Wenjun Zhang. Cross-domain detection via graph-induced prototype alignment. In *CVPR*, pages 12355–12364, 2020.

[72] Hongliang Yan, Yukang Ding, Peihua Li, Qilong Wang, Yong Xu, and Wangmeng Zuo. Mind the class weight bias: Weighted maximum mean discrepancy for unsupervised domain adaptation. In *CVPR*, pages 2272–2281, 2017.

[73] Hantao Yao, Rui Zhang, and Changsheng Xu. Visual-language prompt tuning with knowledge-guided context optimization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6757–6767, 2023.

[74] Hantao Yao, Rui Zhang, and Changsheng Xu. Tcp: Textual-based class-aware prompt tuning for visual-language model. In *CVPR*, pages 23438–23448, 2024.

[75] Jayeon Yoo, Inseop Chung, and Nojun Kwak. Unsupervised domain adaptation for one-stage object detector using offsets to bounding box. In *ECCV*, pages 691–708, 2022.

[76] Xu Zhang, Felix Xinnan Yu, Shih-Fu Chang, and Shengjin Wang. Deep transfer network: Unsupervised domain adaptation. *arXiv preprint arXiv:1503.00591*, 2015.

[77] Yixin Zhang, Zilei Wang, and Yushi Mao. Rpn prototype alignment for domain adaptive object detector. In *CVPR*, pages 12425–12434, 2021.

[78] Liang Zhao and Limin Wang. Task-specific inconsistency alignment for domain adaptive object detection. In *CVPR*, pages 14217–14226, 2022.

[79] Yiwu Zhong, Jianwei Yang, Pengchuan Zhang, Chunyuan Li, Noel Codella, Liunian Harold Li, Luowei Zhou, Xiyang Dai, Lu Yuan, Yin Li, et al. Regionclip: Region-based language-image pretraining. In *CVPR*, pages 16793–16803, 2022.

[80] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Conditional prompt learning for vision-language models. In *CVPR*, pages 16816–16825, 2022.

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

[82] Qianyu Zhou, Qiqi Gu, Jiangmiao Pang, Zhengyang Feng, Guangliang Cheng, Xuequan Lu, Jianping Shi, and Lizhuang Ma. Self-adversarial disentangling for specific domain adaptation. *TPAMI*, 45(7):8954–8968, 2023.

[83] Wenzhang Zhou, Heng Fan, Tiejian Luo, and Libo Zhang. Unsupervised domain adaptive detection with network stability analysis. In *ICCV*, pages 6986–6995, 2023.

[84] Yangguang Zhu, Ping Guo, Haoran Wei, Xin Zhao, and Xiangbin Wu. Disentangled discriminator for unsupervised domain adaptation on object detection. In *IROS*, pages 5685–5691, 2023.

[85] Yaoming Zhu, Jiangtao Feng, Chengqi Zhao, Mingxuan Wang, and Lei Li. Serial or parallel? plug-able adapter for multilingual machine translation. *arXiv preprint arXiv:2104.08154*, 2021.

6 Appendix

6.1 Datasets

Cross-Weather Cityscapes [9] contains diverse street scenes captured by a mobile camera in daylight. The regular partition consists of 2,975 training and 500 validation images annotated with eight classes. Foggy Cityscapes [54] simulates three distinct densities of fog on Cityscapes, containing 8,925 training images and 1,500 validation images. A standard configuration for cross-weather adaptation is to take the training set of Cityscapes as the source domain and the training set of foggy Cityscapes as the target domain, evaluating cross-weather adaptation performance on the 1500-sized validation set in all eight categories.

Cross-FoV KITTI [18] is a crucial dataset for self-driving that includes 7,481 photos annotated with cars. Collected by driving in rural areas and on highways, it provides data with a different Field of View (FoV). To fairly compare with other methods, we migrate KITTI to Cityscapes solely on the car category.

Sim-to-Real The synthetic dataset SIM10k [28] has 10,000 photos from the Grand Theft Auto V video game with labelled bounding boxes in the class car. We follow existing works to perform this sim-to-real adaptation and report the performance on the class car.

Cross-Style Pascal VOC [13] is a widely-used real-world dataset containing 2007 and 2012 subsets, annotated with 20 classes. Clipart [26] is collected from the website with 1000 comical images, providing bounding box annotations with same classes as Pascal VOC. Following the mainstream splitting, we use Pascal VOC 2007 and 2012 train-val split with a total of 16,551 images as the source domain and all Clipart images as the target domain.

6.2 Implementation Details

Following [31], we adapt RegionCLIP(ResNet-50 [22]) with a domain classifier [7] as the baseline. We use the Faster-RCNN [52] as the detector with the default configurations. Following [7], one batch of source images with ground truth and one batch of target domain images are forwarded to the proposed DA-Ada in each iteration to calculate the detection, adversarial and decoupling loss. The hyperparameter $\lambda_{dia}, \lambda_{dita}, \lambda_{dec}$ is set to 0.1, 1.0 and 0.1, respectively. We set the batch size of each domain to 8 and use the SGD optimizer with a warm-up learning rate. Mean Average Precision (mAP) with a threshold of 0.5 is taken as the evaluation metric. All experiments are deployed on 8 Tesla V100 GPUs.

6.3 Comparison to SOTA methods

We present representative state-of-the-art DAOD approaches for comparison, including feature alignment and semi-supervised learning methods.

Cross-Weather Adaptation Scenario Table 11 (C→F) illustrates that the proposed DA-Ada surpasses SOTA DA-Pro [31] by a remarkable margin of 2.6%, achieving the highest mAP over eight classes of 58.5%. Compared with existing methods, our method significantly improves seven categories (*i.e.* person, rider, car, truck, bus, train, and bicycle) ranging from 0.4% to 5.3%. Compared with the SOTA decoupling method D^2-UDA [84], the DA-Ada attains an improvement of 7.9% and promotes $3.0 \sim 10.9\%$ on all categories. The superior performance shows modifying domain-invariant knowledge with domain-specific knowledge could enhance the discriminative capability on the target domain.

Cross-FOV Adaptation Scenario Table 11 (K→C) indicates a noticeable 5.3% improvement on SOTA DA-Pro [31] by the DA-Ada, reaching an astounding peak of 66.7% mAP. As K→C adaptation faces more complicated shape confusion than C→F, it requests higher discriminability of the model. Therefore, the considerable enhancement validates that the proposed method can efficiently improve the discriminability of the visual encoder in new scenarios.

Sim-to-Real Adaptation Scenario We report the experimental results on SIM10k → Cityscapes benchmark in Table 11 (S→C). The proposed DA-Ada achieves the best results of 67.3% mAP, outperforming the previous best entry HT [11] 65.5% with 1.8%. The performance of DA-Ada is

Table 11: Comparison (%) with existing methods on Cross-Weather adaptation Cityscapes→Foggy Cityscapes (C→F), Cross-Fov adaptation KITTI→Cityscapes (K→C) and Sim-to-Real adaptation SIM10K→Cityscapes (S→C).

	C→F									K→C	S→ C
Methods	Person	Rider	Car	Truck	Bus	Train	Motor	Bicycle	mAP	mAP	mAP
DA-Faster [7]	29.2	40.4	43.4	19.7	38.3	28.5	23.7	32.7	32.0	41.9	38.2
VDD [68]	33.4	44.0	51.7	33.9	52.0	34.7	34.2	36.8	40.0	-	-
DSS [65]	42.9	51.2	53.6	33.6	49.2	18.9	36.2	41.8	40.9	42.7	44.5
MeGA [62]	37.7	49.0	52.4	25.4	49.2	46.9	34.5	39.0	41.8	43.0	44.8
SCAN [36]	41.7	43.9	57.3	28.7	48.6	48.7	31.0	37.3	42.1	45.8	52.6
TIA [78]	52.1	38.1	49.7	37.7	34.8	46.3	48.6	31.1	42.3	44.0	-
DDF [42]	37.6	45.5	56.1	30.7	50.4	47.0	31.1	39.8	42.3	46.0	44.3
SIGMA [37]	44.0	43.9	60.3	31.6	50.4	51.5	31.7	40.6	44.2	45.8	53.7
SIGMA++ [38]	46.4	45.1	61.0	32.1	52.2	44.6	34.8	39.9	44.5	49.5	57.7
CIGAR [44]	46.1	47.3	62.1	27.8	56.6	44.3	33.7	41.3	44.9	48.5	58.5
SAD [82]	38.3	47.2	58.8	34.9	57.7	48.3	35.7	42.0	45.2	-	49.2
OADA [75]	47.8	46.5	62.9	32.1	48.5	50.9	34.3	39.8	45.4	47.8	59.2
CSDA [16]	46.6	46.3	63.1	28.1	56.3	53.7	33.1	39.1	45.8	48.6	57.8
HT [11]	52.1	55.8	67.5	32.7	55.9	49.1	40.1	50.3	50.4	60.3	65.5
D^2-UDA [84]	46.9	53.3	64.5	38.9	61.0	48.5	42.6	54.2	50.6	60.3	58.1
AT [39]	56.3	51.9	64.2	38.5	45.5	55.1	**54.3**	35.0	50.9	-	-
NSA-UDA [83]	50.3	60.1	67.7	37.4	57.4	46.9	47.3	54.3	52.7	55.6	56.3
DA-Pro [31]	55.4	62.9	70.9	40.3	63.4	54.0	42.3	58.0	55.9	61.4	62.9
DA-Ada(Ours)	**57.8**	**65.1**	**71.3**	**43.1**	**64.0**	**58.6**	48.8	**58.7**	**58.5**(±0.2)	**66.7**(±0.3)	**67.3**(±0.2)

Table 12: Comparison (%) with existing methods on Cross-Style adaptation task Pascal VOC→Clipart

Methods	Aero	Bike	Bird	Boat	Bottle	Bus	Car	Cat	Chair	Cow	Table	Dog	Horse	Motor	Person	Plant	Sheep	Sofa	Train	Tv	mAP
UaDAN [20]	35.0	73.7	41.0	24.4	21.3	69.8	53.5	2.3	34.2	61.2	31.0	**29.5**	47.9	63.6	62.2	61.3	13.9	7.6	48.6	23.9	40.2
TFD [63]	27.9	64.8	28.4	29.5	25.7	64.2	47.7	13.5	47.5	50.9	50.8	21.3	33.9	60.2	65.6	42.5	15.1	40.5	45.5	48.6	41.2
DBGL [3]	28.5	52.3	34.3	32.8	38.6	66.4	38.2	25.3	39.9	47.4	23.9	17.9	38.9	78.3	61.2	51.7	26.2	28.9	56.8	44.5	41.6
IIPD [67]	41.5	52.7	34.5	28.1	43.7	58.5	41.8	15.3	40.1	54.4	26.7	28.5	37.7	75.4	63.7	48.7	16.5	30.8	54.5	48.7	42.1
FGRR [4]	30.8	52.1	35.1	32.4	42.2	62.8	42.6	21.4	42.8	58.6	33.5	20.8	37.2	81.4	66.2	50.3	21.5	29.3	**58.2**	47.0	43.3
UMT [10]	39.6	59.1	32.4	35.0	45.1	61.9	48.4	7.5	46.0	**67.6**	21.4	**29.5**	48.2	75.9	70.5	**56.7**	25.9	28.9	39.4	43.6	44.1
SIGMA [37]	40.1	55.4	37.4	31.1	54.9	54.3	46.6	23.0	44.7	65.6	23.0	22.0	42.8	55.6	67.2	55.2	32.9	**40.8**	45.0	58.6	44.5
ATMT [35]	37.5	63.4	37.9	29.8	45.1	62.7	41.2	19.5	43.7	57.4	22.9	25.3	39.6	87.1	70.9	50.6	29.1	32.2	58.4	50.5	45.2
CIGAR [44]	35.2	55.0	39.2	30.7	60.1	58.1	46.9	31.8	47.0	61.0	21.8	26.7	44.6	52.4	68.5	54.4	31.3	38.8	56.5	63.5	46.2
TIA [78]	42.2	66.0	36.9	37.3	43.7	**71.8**	49.7	18.2	44.9	58.9	18.2	29.1	40.7	**87.8**	67.4	49.7	27.4	27.8	57.1	50.6	46.3
SIGMA++ [38]	36.3	54.6	40.1	31.6	58.0	60.4	46.2	**33.6**	44.4	66.2	25.7	25.3	44.4	58.8	64.8	55.4	36.2	38.6	54.1	**59.3**	46.7
CMT [60]	39.8	56.3	38.7	39.7	**60.4**	35.0	**56.0**	7.1	**60.1**	60.4	**35.8**	28.1	**67.8**	84.5	**80.1**	55.5	20.3	32.8	42.3	38.2	47.0
DA-Ada(Ours)	**42.3**	**75.1**	**48.9**	**45.9**	49.0	**71.8**	55.6	15.4	50.7	56.6	19.9	20.6	61.3	80.7	73.0	29.2	**37.5**	21.5	52.5	52.9	**48.0**(±0.1)

superior in the difficult adaptation task, which further demonstrates that our strategy is robust not only in appearance but also in more complex semantics adaptation tasks.

Cross-Style Adaptation Scenario Additionally, we assess DA-Ada on the more challenging Cross-Style adaptation, where the semantic hierarchy has a broader domain gap. DA-Ada peaks with 48.0%, outperforming all the SOTA methods presented in Table 12. It demonstrates that injecting cross-domain information into the visual encoder could benefit the adaptation. Especially, DA-Ada exceeds all the compared methods on six categories (aeroplane, bike, bird, boat, bus, and sheep), which verifies the method is effective under challenging domain shifts and in multi-class problem scenarios.

6.4 Sensitivity on $\mathcal{L}_{dia}$

Table 13: Sensitivity to hyper-parameters of initialization of λ_{dia}.

Cityscapes→ FoggyCityscapes						
λ_{dia}	0.01	0.05	0.1	0.5	1.0	10.0
mAP	57.1	57.8	58.5	58.1	58.0	53.4

To select hyper-parameters for the adversarial loss in DIA, we perform experiments of different choices of the weight value λ_{dia}. We conduct the experiment on DA-Ada on

Cityscapes→FoggyCityscapes adaptation scenarios, as shown in Table 13. Initialized with 0.01, the DIA suffers from insufficient learning of domain-invariant knowledge, only attaining 57.1% mAP. When initialized with $0.05 \sim 1.0$, the performance of DA-Ada is similar and achieves the best of 58.5% with $\lambda_{dia} = 1.0$. Increasing the λ_{dia} to 10.0 sufferers significant performance degradation. We attribute this to the model focusing too much on alignment rather than detection.

6.5 Sensitivity on λ_{dita}

Table 14: Sensitivity to hyper-parameters of initialization of $\mathcal{L}_{dita}$.

Cityscapes→ FoggyCityscapes					
λ_{dita}	0.1	0.5	1.0	5.0	10.0
mAP	58.1	58.2	58.5	57.9	56.9

We also explicitly study the sensitivity of weight value λ_{dita} for the adversarial loss in visual-guided domain prompt, as shown in Table 14. As the weight value increases, the performance peaks with $\lambda_{dita} = 1.0$ and then appears to decline. Since the hand-crafted token "A photo of [CLASS]" provides a solid prior, the learnable prompt is better initialized in the early stages of training. Therefore, the proposed model is more robust to the λ_{dita} compared to λ_{dia}.

6.6 Sensitivity on λ_{dec}

Table 15: Sensitivity to hyper-parameters of initialization of $\mathcal{L}_{dec}$.

Cityscapes→ FoggyCityscapes					
λ_{dec}	0.01	0.05	0.1	0.5	1.0
mAP	57.5	57.9	58.5	58.3	58.4

We also evaluate the sensitivity of weight value λ_{dec} for the decouple loss between DIA and DSA, as shown in Table 15. As the weight value increases, the performance rises rapidly until $\lambda_{dec} = 1.0$ and then declines smoothly. $\mathcal{L}_{dec}$ decouples domain-invariant and domain-specific knowledge by driving DIA to be orthogonal to the features extracted by DSA. Therefore, applying the decoupling loss with the same scale as the adversarial loss can optimize the goal relatively stably.

6.7 Ablation for Multi-scale Down-projector $\mathcal{C}^D$

Table 16: Ablation for Different Resolution in Multi-scale Down-projector $\mathcal{C}^D$.

Cityscapes→ FoggyCityscapes					
Resolutions					mAP
1	$\frac{1}{2}$	$\frac{1}{4}$	$\frac{1}{8}$	$\frac{1}{16}$	
✓					57.6
✓	✓				57.8
✓	✓	✓			58.3
✓	✓	✓	✓		58.5
✓	✓	✓	✓	✓	58.2

Table 16 shows the impact of the different resolutions in multi-scale down-projector $\mathcal{C}^D$. The results indicate that while introducing various resolutions contributes to modeling multi-scale domain knowledge, inappropriate receptive fields may harm feature extraction performance. Concretely, we observed that the mAP on the target domain (Foggy Cityscapes) peaks when the number of down-projectors M is set to 4, and the scaling ratios are $\{1, \frac{1}{2}, \frac{1}{4}, \frac{1}{8}\}$, respectively. And further applying $\frac{1}{16}$ to $\mathcal{C}^D$ results in slightly performance degradation. These experiments suggest that

applying a single convolution or introducing excessive distinct resolutions is unsuitable for learning domain knowledge, and the choice of resolution requires consideration of the difference in scale between the source and target domains.

6.8 Image or Instance-level Visual-guided Textual Adapter?

We explore whether to apply the visual-guided textual adapter at the image or instance levels. In DITA and DSTA, we replace the proposal embedding with the entire image as visual input, and it achieves 57.8%, suffering a performance drop of 0.7%. This indicates that instance-level alignment avoids the influence of background on learning domain knowledge in the foreground.

6.9 Evaluation on multiple Baselines

Table 17: Results of multiple Baselines on C→F adaptation.

Baseline	w/o DA-Ada	with DA-Ada	Gains
DSS	40.9	48.1	+7.2
CLIP+Faster-RCNN	42.8	52.6	+9.8

To properly evaluate the method, we introduce Da-Ada to two weaker baselines in 17. We inject DA-Ada into DSS and attain 7.2% improvement on mAP, showing great efficiency in feature-alignment methods. For further validation, we first adapt the classification model CLIP to Faster-RCNN to build a vanilla VLM detector with 42.8% mAP. Then we freeze the backbone and attach DA-Ada to the detector, achieving 52.6% mAP with an improvement of 9.8%. Experiments show that even with weak baselines, the proposed DA-Ada shows competitiveness to SOTA methods.

6.10 Performance and Computational Overhead

Table 18: Comparison on performance and computational overhead on C→F adaptation.

Method	mAP	Inference time(s)/iter	Training time(s)/iter	Total iter	Mem usg.(MB)
Global Fine-tune	53.6	0.40	2.67	25000	12977
DA-Pro	54.6	0.40	1.47	1000	4034
DA-Ada	58.5	0.42	1.61	2500	6046

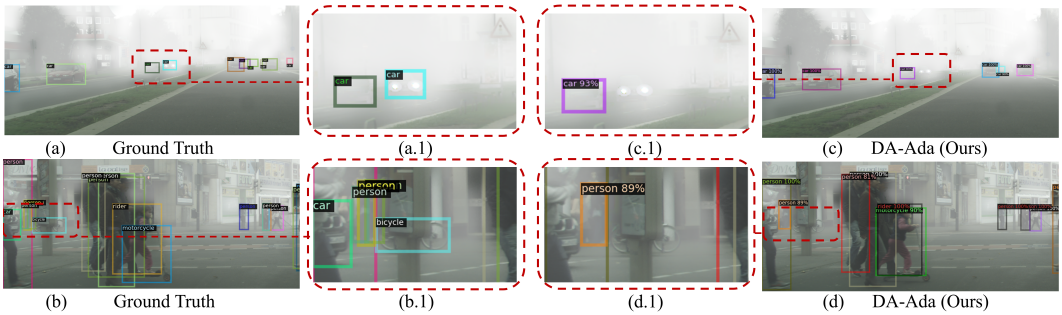
To verify the effectiveness of DA-Ada, we compare the performance and computational overhead with global fine-tune and DA-Pro on C→F adaptation. We initial the three methods with the same VLM backbone. Global fine-tuning has the largest time and memory overhead but only achieves the lowest performance, indicating the limitations of traditional DAOD methods in optimizing VLM. Compared with global fine-tuning, DA-Pro significantly reduces time and memory overhead while improving performance. Furthermore, DA-Ada significantly improves mAP with 4.9% while only using 6% of the time and 47% of the memory, showing great efficiency in adapting cross-domain information to VLM.

6.11 Failure Cases

We provide some examples of failure cases on the Cross-Weather adaptation scenario in Fig 6. We visualize the ground truth (a)(b) and the detection boxes of DA-Ada (c)(d). In (c.1), DA-Ada misses the car with its headlights on in the fog. Since the source data Cityscapes is collected on sunny days, few cars turned on their lights in the training set. Therefore, DA-Ada missed such out-of-distribution data. In (d.1), DA-Ada misses the bicycle and person blocked by other foreground objects. Since occlusion causes great damage to semantics, this type of missed detection is widely seen in object detection methods.

6.12 Limitation

Though effective, the proposed DA-Ada is specially designed for the domain adaptive object detection task, where a labelled source domain and a unlabelled target domain are needed. Currently, the



(a) Ground Truth (a.1) (c.1) (c) DA-Ada (Ours)

(b) Ground Truth (b.1) (d.1) (d) DA-Ada (Ours)

Figure 6: Examples of failure cases on the Cross-Weather adaptation scenario. We visualize the ground truth (a)(b) and the detection boxes of DA-Ada (c)(d). (a.1)(b.1)(c.1)(d.1) are zoomed from corresponding region for better view.

method cannot deal with the setting of multiple source domains or no target domain. We plan to resolve these problems in our future research.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: We state the contribution in the abstract and introduction, and provide experimental results in the last paragraph introduction.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We have discussed the limitation of the proposed method in Sec. 6.12 of the Appendix.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [NA]

Justification: The paper does not include theoretical results.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: We provide source code in supplement material.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: We provide source code and instructions in supplement material.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: We detailed the setting of datasets and the implementation in Sec. 6.1 and 6.2 of the Appendix.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: We provide error bars in Table 11, 12. The error bars are captured by multiple running with given experimental conditions.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).

- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We provide the compute works in Sec. 6.2, and discuss the computation overhead in Sec. 6.10 of the Appendix.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The research conforms with the NeurIPS Code of Ethics.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [NA]

    Justification: There is no societal impact of the work performed.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.

- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: This paper poses no such risks.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We cite the original paper that produced the code package or dataset.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.

- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification: This paper does not release new assets.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: This paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: This paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.