Triple Feature Disentanglement for One-Stage Adaptive Object Detection

Haoan Wang[1], Shilong Jia[1], Tieyong Zeng[2], Guixu Zhang[1], Zhi Li[1]*

[1]School of Computer Science and Technology, East China Normal University, Shanghai, China
[2]Department of Mathematics, The Chinese University of Hong Kong, Shatin, NT, Hong Kong
{haw, 51255901119}@stu.ecnu.edu.cn, zeng@math.cuhk.edu.hk, {gxzhang, zli}@cs.ecnu.edu.cn

Abstract

In recent advancements concerning Domain Adaptive Object Detection (DAOD), unsupervised domain adaptation techniques have proven instrumental. These methods enable enhanced detection capabilities within unlabeled target domains by mitigating distribution differences between source and target domains. A subset of DAOD methods employs disentangled learning to segregate Domain-Specific Representations (DSR) and Domain-Invariant Representations (DIR), with ultimate predictions relying on the latter. Current practices in disentanglement, however, often lead to DIR containing residual domain-specific information. To address this, we introduce the Multi-level Disentanglement Module (MDM) that progressively disentangles DIR, enhancing comprehensive disentanglement. Additionally, our proposed Cyclic Disentanglement Module (CDM) facilitates DSR separation. To refine the process further, we employ the Categorical Features Disentanglement Module (CFDM) to isolate DIR and DSR, coupled with category alignment across scales for improved source-target domain alignment. Given its practical suitability, our model is constructed upon the foundational framework of the Single Shot MultiBox Detector (SSD), which is a one-stage object detection approach. Experimental validation highlights the effectiveness of our method, demonstrating its state-of-the-art performance across three benchmark datasets.

Introduction

Due to the availability of numerous annotated datasets, object detection techniques based on deep learning have made significant advancements (Zhu et al. 2021; Liu et al. 2021b; He et al. 2022). Lately, models that rely on full supervision have been criticized for their apparent deficiency in terms of generalization ability (Liu et al. 2022). Firstly, performance drops significantly when applying the detector to a new domain. Besides, on unlabeled datasets, detectors generally hard to yield satisfactory results. Furthermore, the task of labeling datasets requires a significant amount of time and resources.

To address these problems, developing an algorithm that can transfer the knowledge learned from a labeled source domain to another unlabeled target domain becomes critical for object detection (Yao et al. 2021; Li et al. 2022c). Therefore, Unsupervised Domain Adaptation (UDA) is applied to the field of object detection (Zhang et al. 2021a; Liu et al. 2021a; Jin et al. 2021; Zhu et al. 2022) and a novel task called Domain Adaptive Object Detection (DAOD) (Chen et al. 2018) is proposed to narrow the domain shift between two domains. Existing methods (Vs et al. 2021; Zhao and Wang 2022; Xu et al. 2022a) use Gradient Reversal Layers (GRL) (Ganin and Lempitsky 2015) and domain discriminators for this purpose. In this way, the detector is able to achieve acceptable detection performance on the unlabeled target domain. However, the above methods may ignore the influence of domain-private information (Peng et al. 2019; Liu et al. 2022). The inherent instability of the adversarial learning process contributes to the inaccuracy of prediction results in these methods (Arjovsky, Chintala, and Bottou 2017). Furthermore, the features extracted in conventional adversarial learning-based UDA approaches fail to achieve domain-invariant and inevitably retain domain-private factors. When optimizing UDA models using these entangled features, the bias towards the source domain further amplifies the degradation of performance in the target domain (Liu et al. 2022).

For DAOD tasks, it is important to obtain Domain-Invariant Representations (DIR), which can lessen the effect of domain shift (Wu et al. 2021). Hence, traditional ap-

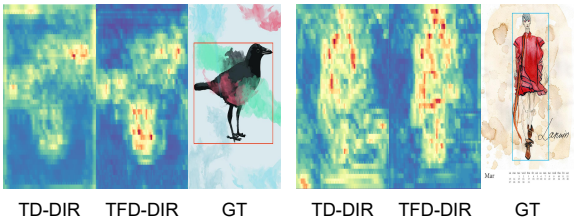

Figure 1: Comparisons of feature maps extracted by Traditional Disentanglement (TD) and our Triple Feature Disentanglement (TFD). The extracted DIR from TD and TFD are represented as 'TD-DIR' and 'TFD-DIR', respectively. The DIR obtained through our TFD contains much less domain-specific information, thereby enabling more emphasis on the objects of interest.

*Corresponding author.

proaches (Wu et al. 2021; Lang et al. 2022) exploit the concept of disentangled learning and aim to separate Domain-Specific Representations (DSR) and DIR. The final classification and regression predictions rely on DIR, which should contain as little domain-private information as possible. However, as shown in Figure 1, these existing methods (Wu et al. 2021; Lang et al. 2022) employ simple disentanglement operation, which cannot effectively separate DIR and DSR (Cai et al. 2019b). The DIR extracted through the simple operation may incorporate domain-private information, ultimately leading to detection errors due to a lack of accurate representation of the region of interest.

Therefore, we propose a novel method named Triple Feature Disentanglement (TFD) in this paper. In TFD, disentanglement is accomplished through three phases. Firstly, we can yield more pure DIR with Multi-level Disentanglement Module (MDM). In addition, we use the Cyclic Disentanglement Module (CDM) to extract DSR and then increase the gap between DIR and the cyclically extracted DSR. It can bring DIR closer to completed disentanglement. In addition, we also design Spatial Pyramid Pooling Attention (SPP Attention) to assist in extracting domain-private information. Furthermore, different categories pose distinct challenges for cross-domain alignment. Targeting this problem, Categorical Features Disentanglement Module (CFDM) is proposed to disentangle category features and align the same categories. Moreover, we observe that the multi-scale strategy has rarely been recognized in the past. However, the extracted feature distributions at various scales are most likely to be distinct (Chen et al. 2021c). Therefore, category-level alignment is applied to all scales.

Additionally, the majority of the current approaches are based on the two-stage detector Faster R-CNN (Ren et al. 2015). However, it primarily relies on ROI-based instance-level features and the region proposal mechanism (Chen et al. 2021b), which require a significant amount of computational time. Moreover, the DAOD techniques that rely on Faster R-CNN are tailored specifically to suit region proposal network (RPN) (Cai et al. 2019a; Zhu et al. 2019), which makes it almost impossible to apply these methods to other detectors. As we all know, one-stage object detection algorithms are faster and better suited for practical application (Sultana, Sufian, and Dutta 2020; Jiang et al. 2020). The absence of a special structure in SSD (Liu et al. 2016) enables easy transfer of methods designed around it to different detectors. Therefore, we use the one-stage detector SSD as the baseline for our domain adaptive detector instead of two-stage ones.

The following is a summary of the main contributions of this paper:

- In contrast to previous straightforward disentanglement methods, our technique employs disentanglement across three distinct phases. The initial phase involves a multi-level Disentanglement Module (MDM). We propose a Re-disentanglement Block (RB), building upon the initial disentanglement process. Experimental results demonstrate the significant role played by the RB in enhancing performance.
- To facilitate the extraction of DSR, we propose the Spatial Pyramid Pooling (SPP) Attention mechanism in CDM. SPP Attention effectively preserves the private information within DSR, which proves beneficial for maintaining a distinct separation from the DIR during the subsequent stages of the training process.
- The Categorical Features Disentanglement Module (CFDM) has been presented as a way of separating category-level features and is divided into two parts. Part 1 is utilized to ensure that category features produced by DIR have little domain-private information. Part 2 promotes the alignment of inter-domain features corresponding to the same category while driving apart intra-domain features corresponding to different categories.
- Our method showcases superior performance compared to existing methods, as validated by experiments on three public datasets, e.g., improving the adaptation performance on Pascal VOC to Watercolor2K from 53.5% mAP to 55.0% mAP.

Related Work

Unsupervised Domain Adaptation

The objective of Unsupervised Domain Adaptation (UDA) is to transfer the knowledge from a labeled source domain to an unlabeled target domain, thus reducing the time and resources required for labeling. UDA is commonly employed in computer vision tasks, including but not limited to classification (Yu, Zhai, and Zhang 2022; Yin et al. 2022) and segmentation (Li et al. 2022a; Zhang et al. 2021b). Yu et al. (Yu, Zhai, and Zhang 2022) proposed a two-stage adaptive classification framework, wherein the first stage aimed to reduce misalignment in the methods used for learning invariant representations, while the second stage focused on encoding the semantic structure of unlabeled target data. Li et al. (Li et al. 2022a) adopted UDA in the field of semantic segmentation and argued that the information in the target domain should be thoroughly explored from the views of both boundaries and features. Compared with the above-mentioned tasks, object detection poses a greater challenge, requiring the optimization of two subtasks (classification and localization) simultaneously (Zhao and Wang 2022). Our model tackles this intricate problem from a fresh perspective by using disentangled learning to achieve cross-domain alignment.

Disentangled Learning

Disentangled learning has many applications in the field of domain adaptation such as domain adaptive segmentation (Zhao et al. 2023; Yang et al. 2019) and domain adaptive classification (Zhou et al. 2020; Wu et al. 2022). In (Yang et al. 2019), disentangled learning was employed to extract domain-invariant representations for cross-domain liver segmentation. Zhou et al. (Zhou et al. 2020) first disentangled DIR and DSR and created prototypes accordingly, then reconstructed features obtained from prototypes were utilized to achieve classification results. Disentangled learning has been found promising in the field of UDA. However, its application in DAOD is still largely unexplored, which motivates us to delve further into this area of research.

Object Detection

Object detection can be broadly classified into two-stage and one-stage methods. Faster R-CNN (Ren et al. 2015) is a well-known two-stage approach, which uses RPN to generate coarse object proposals and then feeds the proposals into a classification module. Although Faster R-CNN has been extensively utilized in the field of DAOD, its practical application has been hindered by its limited detection speed (Kim, Sung, and Park 2020). One-stage methods, on the other hand, were proposed without the generation of region proposals. SSD (Liu et al. 2016), as a typical one-stage approach, generates multi-scale feature maps to obtain detection results. This method has gained widespread popularity due to its superior detection speed and performance. (Araki et al. 2020; Chen et al. 2021b,a).

Disentangled Learning for DAOD

Disentangled learning is an active research area, until recently, its applications have extended to the realm of Domain Adaptive Object Detection (DAOD) (Wu et al. 2021; Liu et al. 2022; Lang et al. 2022). Nonetheless, this is a relatively new direction of research, there are only a few studies available on the topic. VDD (Wu et al. 2021) focused on feature disentanglement and employed a novel method based on vector decomposition that encouraged DIR to contain more domain-invariant information. Liu et al. (Liu et al. 2022) developed DDF that facilitated feature disentanglement at the global and local stages, respectively. Lang et al. (Lang et al. 2022) proposed IDF, which used non-adversarial domain discriminators, dual attention mechanisms, and selective feature perception to achieve feature disentanglement. Despite its potential benefits, disentanglement learning remains overlooked in the DAOD field, with most existing approaches relying on simple disentanglement. In contrast to conventional approaches, our method presents a novel strategy for achieving disentanglement via three distinct mechanisms, with the final goal of minimizing domain-private information contained within DIR.

Method

Before introducing our approach, we review the problem description. We are given N^s labeled images $\mathcal{D}^s = \{(x_i^s, y_i^s)\}_{i=1}^{N^s}$ in source domain, where x_i^s and y_i^s represent the image and annotation, respectively. There is a target domain $\mathcal{D}^t = \{x_j^t\}_{j=1}^{N^t}$ of N^t unlabeled samples. $\mathcal{D}^s$ and $\mathcal{D}^t$ are derived from distinct data distributions, but they share the same set of classes. The ultimate objective of DAOD is to design a domain-invariant detector and transfer knowledge from $\mathcal{D}^s$ to $\mathcal{D}^t$.

In this section, we introduce the Triple Feature Disentanglement (TFD). We will first present the framework of TFD and then describe each component of the network.

Network Architecture

As shown in Figure 2, TFD consists of three components, namely, Multi-level Disentanglement Module (MDM), Cyclic Disentanglement Module (CDM), and Categorical

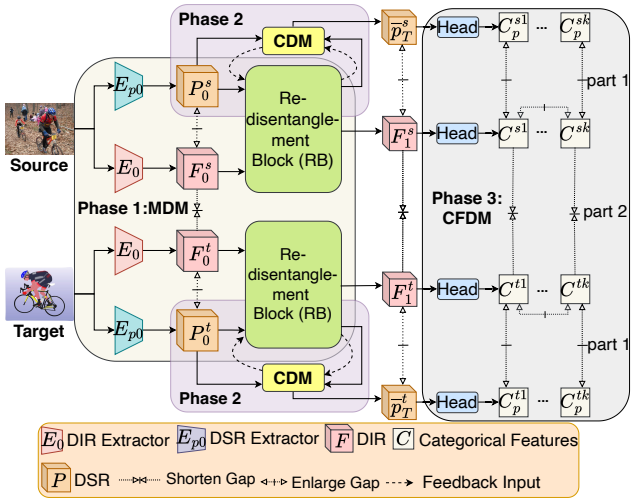

Figure 2: The network architecture of our proposed TFD. T is the number of time steps in CDM and C_p denotes the categorical feature that comes from DSR. The process of disentanglement is carried out in three phases, beginning with multi-level disentanglement achieved through MDM, followed by cyclic disentanglement as the second phase, and ending with category-level disentanglement as the final phase.

Features Disentanglement Module (CFDM). DSR and DIR are separated during the feature extraction process using MDM. CDM re-inputs DSR into an extractor and further captures domain-private features. CFDM aims to increase the gap between distinct category features and align the same category features of two domains. Furthermore, the CFDM approach also accomplishes disentanglement at the category level, thereby increasing the distribution disparity between DIR and DSR even more.

Multi-level Disentanglement Module

Traditional disentanglement methods (Wu et al. 2021; Liu et al. 2022) are performed in a single step, and the extracted DIR may be mixed with domain-private features (Cai et al. 2019b), resulting in suboptimal detector performance. Therefore, we creatively suggest MDM, where the disentanglement procedure is broken down into two steps.

In the initial step, the image is first sent to E_{p0} and E_0, which represent the DSR extractor and the DIR extractor, respectively. Then we can gain feature maps P_0 and F_0, which are DSR and DIR. Nevertheless, we believe that F_0 still contains domain-private features and is not yet fully disentangled. Therefore, we design the Re-disentanglement Block (RB), which will further extract features and make F_0 almost completely disentangled. As depicted in Figure 3, we design a new DSR extractor E_{p1} for RB. F_0 and P_0 produced in the first step are passed to E_1 and E_{p1} in order to obtain F_1, P_1, F_2, and P_2. Among them, F_1 represents the DIR, P_2 is the DSR, and P_1 and F_2 denote domain-private features contained in F_0 and domain-invariant features con-

tained in P_0, respectively.

$$P_0 = E_{p0}(x), P_1 = E_{p1}(F_0), P_2 = E_{p1}(P_0), \quad (1)$$
$$F_0 = E_0(x), F_1 = E_1(F_0), F_2 = E_1(P_0), \quad (2)$$

where F and P represent DIR and DSR respectively.

To further drive DIR and DSR apart, we introduce an orthogonal loss (Wu et al. 2021) with the following formula:

$$\mathcal{L}_{ORT} = \frac{1}{4N}\sum_{i=0}^{3}\sum_{j=1}^{N}\left|\sum_{k=1}^{c}\frac{A_{ijk}}{\|A_{ijk}\|_2^2}\odot\frac{B_{ijk}}{\|B_{ijk}\|_2^2}\right|, \quad (3)$$

where $A = [P_0, P_1, P_2, P_2]$, $B = [F_0, F_1, F_2, F_1]$. Besides, N, c, $\|\cdot\|_2^2$, $|\cdot|$ and $\odot$ represent the number of samples, the number of channels, L2- norm, the absolute value operation, and the element-wise product, respectively. For the sake of conciseness, we omit superscripts s and t throughout the paper for these equations which are applied to both the source and target domains, and the average of the corresponding two losses is used.

Furthermore, we measure the distance between DIR obtained from the source and target domains using the Euclidean distance. In this way, we can align the two domains. As shown in Figure 2, we minimize the gaps between feature maps F_0^s and F_0^t, as well as F_1^s and F_1^t.

$$\mathcal{L}_{ED} = \frac{1}{2N}\sum_{i=0}^{1}\sum_{j=1}^{N}\Phi\left(\left\|F_{ij}^s\right\|_2^2, \left\|F_{ij}^t\right\|_2^2\right), \quad (4)$$

where $\Phi(a, b)$ is the Euclidean distance.

To further narrow the gap between DIR obtained from source and target domains, we introduce domain discriminators and GRL (Ganin and Lempitsky 2015) for adversarial learning. Domain discriminators are employed to determine whether it belongs to the source domain or the target domain. By employing domain discriminators, we can effectively reduce the amount of domain-specific information contained in DIR, while mitigating any potential impact from variations in feature batch size (Liu et al. 2022). GRL operates by inverting the gradient during the backpropagation process, establishing an adversarial interplay between the DIR extractor and the discriminator. In this dynamic, the discriminator's objective is to precisely discern where the DIR come from, whereas the extractor aims to obfuscate this information. Through the utilization of the local domain discriminator D_l for F_0 and the global domain discriminator D_g for F_1, we derive the corresponding $\mathcal{L}_{LOC}$ and $\mathcal{L}_{GLB}$ (Saito et al. 2019). Furthermore, the adversarial loss is $\mathcal{L}_{ADV} = \mathcal{L}_{LOC} + \alpha\mathcal{L}_{GLB}$, where α is set to 0.1 in our experiments. In addition, DSR such as P_0 and P_2 are also fed into D_l and D_g (Liu et al. 2022). But GRL is not used and there is no adversarial training. We want the discriminator to be able to accurately determine which domain P_0 and P_1 belong to, then they possess obvious domain-private information. The loss of the discriminator is the same as $\mathcal{L}_{ADV}$, which is named $\mathcal{L}_{DIS}$. In the end, the loss of the proposed MDM can be expressed as: $\mathcal{L}_{\text{MDM}} = \mathcal{L}_{\text{ORT}} + \mathcal{L}_{ED} + \mathcal{L}_{ADV} + \mathcal{L}_{DIS}$.

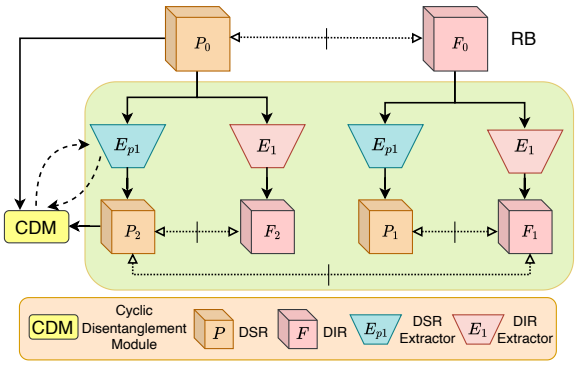

Figure 3: The network architecture of RB.

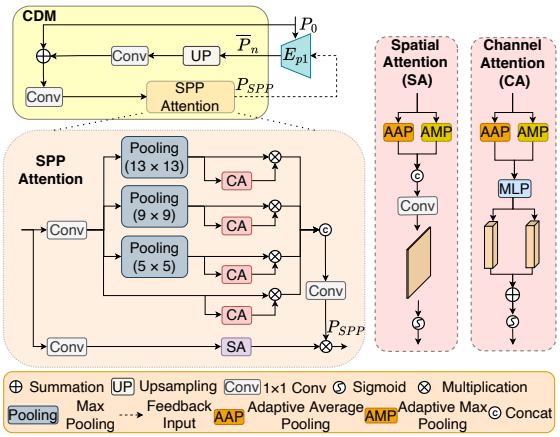

Figure 4: The network architecture of CDM. T is the number of time steps. When time step $n = 1$, P_0 is the input to E_{p1}, $\overline{P}_n$ is its output. But when $n \geq 2$, the input of E_{p1} comes from feedback features P_{SPP} rather than P_0. P_{SPP} represents the output of SPP Attention.

Cyclic Disentanglement Module

Based on the MDM approach, we have developed a new module named the Cyclic Disentanglement Module (CDM) that plays a critical role in the more efficient extraction of DSR by E_{p1}, thereby approaching complete disentanglement. With this module, we can cyclically extract the DSR, which results in obtaining DSR with less domain-invariant information in each cycle. By enlarging the gap between the DSR and DIR, DIR will contain less domain-private information. During testing, only the final prediction results based on DIR are required, and DSR extraction is not necessary, which implies that the cyclic extraction of DSR during training does not increase inference time.

As illustrated in Figure 4, when P_0 is fed into E_{p1}, $\overline{P}_n$ can be obtained, where n is the current time step. Firstly, $\overline{P}_n$ is sent to an upsampling layer and a 1×1 convolution, then added to P_0. Then, it is sent through another 1×1 convolution. To concentrate more on DSR, we develop a block named SPP Attention based on Spatial Pyramid Pooling (SPP) (He et al. 2015). The feature maps generated by the spatial pyramid structure contain more detailed structural

information and are capable of retaining the spatial characteristics within each channel (Ma et al. 2021). As shown in Figure 4, we use multiple max-pooling layers to eliminate redundant spatial information in the feature maps. This is followed by applying the Channel Attention (Woo et al. 2018) to improve the representation power of networks. The feature maps produced by the max-pooling layers are concatenated, and the feature map is subsequently scaled by the spatial attention map, which is derived from the initial feature map, thereby generating the ultimate output denoted as P_{SPP} which is sent back to the E_{p1} for the next cycle.

It is worth noting that the input of E_{p1} comes from feedback features P_{SPP} rather than P_0 when $n \geq 2$. But P_0 is still the input of subsequent skip connections.

To make F_1 contain less domain-private information, we use cosine similarity to enlarge the gap between F_1 and $\overline{P}_n$. The formula is as follows:

$$\mathcal{L}_{CDM} = \frac{1}{\mathrm{TN}} \sum_{n=1}^{T} \sum_{j=1}^{N} sim\left((F_1)_j , \overline{P}_{nj} \right), \tag{5}$$

where T is the number of time steps and $sim(a, b)$ denotes the cosine similarity.

Categorical Features Disentanglement Module

Class-agnostic cross-domain feature alignment only aligns global distributions of two domains. As a result, misalignment of classes will emerge, jeopardizing the ultimate result (Xu et al. 2022b). Due to this reason, we propose a Categorical Features Disentanglement Module (CFDM). CFDM can be broken down into two parts. To achieve category-level disentanglement within each domain, the first part of the module is utilized. Part 2 promotes the alignment of inter-domain features corresponding to the same category while driving apart intra-domain features corresponding to different categories. In the meantime, different scales imply distinct distributions, and their domain offsets are not the same (Chen et al. 2021c). Considering this, we believe that scale is an essential factor in solving the cross-domain alignment issue. Therefore, CFDM is utilized to implement the multi-scale alignment.

As shown in Figure 2, $\overline{P}_T$ is obtained by CDM cyclic extraction and F_1 is obtained by MDM. The aforementioned outputs are subsequently fed into the detection head for the purpose of procuring categorization and regression predictions. Afterward, the formula (6) is employed to derive the category feature information C^k (Chen et al. 2021b).

$$C^k = \frac{1}{N_c} \sum_{i=1}^{N} \sum_{m=1}^{W \times H} \hat{y} \cdot (F_1)_{im}, \tag{6}$$

where N is the number of samples and N_c denotes the number of categories in the dataset. $\hat{y} \in \{0, 1\}$ indicates whether the current pixel is predicted as category k. It is worth noting that the y value originates from the detection head. Once C^k is obtained, the category-level disentanglement is implemented in the following two parts.

In part 1, to enhance the distinction between identical categories in DIR and DSR, we employ cosine similarity for the Category-level Disentanglement (CD) acquired by $\overline{P}_T$ and F_1.

$$\mathcal{L}_{CD} = \frac{1}{LN_c} \sum_{i=1}^{L} \sum_{j=1}^{N_c} \mathrm{sim}\left(C_i^j, (C_p)_i^j \right), \tag{7}$$

where L is the number of feature maps.

In part 2, two distinct loss functions are employed to align categorical features acquired by DIR in the source and target domains. At each scale, we use Jensen-Shannon (JS) divergence to narrow the gap between the same classes in source and target domains. The formula is as follows:

$$\begin{aligned} \mathcal{L}_{JS} = \frac{1}{2LN_c} \sum_{i=1}^{L} \sum_{j=1}^{N_c} \Big[& D_{KL}\left((C^s)_i^j , (C^t)_i^j \right) \\ & + D_{KL}\left((C^t)_i^j , (C^s)_i^j \right) \Big], \end{aligned} \tag{8}$$

$$D_{KL}(a, b) = \frac{1}{N_e} \sum_{i=1}^{N_e} a_i \left(\log \frac{2a_i}{a_i + b_i} \right), \tag{9}$$

where N_e represents the number of elements in a and b. Meanwhile, we not only consider gaps between features of the same categories but also gaps between features of different categories. Increasing the separation between various categories helps minimize the inclusion of information from other categories within a specific category. This is advantageous for cross-domain alignment of categories in preceding steps, ultimately aiding in mitigating the domain shift between the two domains. At each scale, cosine similarity is employed to enlarge the distances of various class features. The loss function is as follows:

$$\mathcal{L}_{COS} = \frac{1}{2LN_c} \sum_{i=1}^{L} \sum_{j=1}^{N_c} \left[\sum_{k=j+1}^{N_c} sim\left(C_i^j, C_i^k \right) \right]. \tag{10}$$

Finally, the objective of CFDM is formulated as: $\mathcal{L}_{CFDM} = \mathcal{L}_{JS} + \gamma_1 \mathcal{L}_{CD} + \gamma_2 \mathcal{L}_{COS}$, where γ_1 and γ_2 is set to 0.1.

Overall Loss

The detection loss of the source domain is denoted as $\mathcal{L}_{DET}$. Besides $\mathcal{L}_{DET}$, the overall loss consists of losses from three other modules, namely MDM, CDM, and CFDM:

$$\mathcal{L}_{TFD} = \mathcal{L}_{DET} + \lambda_1 \mathcal{L}_{MDM} + \lambda_2 \mathcal{L}_{CDM} + \lambda_3 \mathcal{L}_{CFDM}, \tag{11}$$

where λ_1, λ_2 and λ_3 are hyper-parameters.

Experiments

Datasets and Implementation Details

Following the setup in (Kim et al. 2019; Chen et al. 2021b), our proposed approach is evaluated on four datasets. The Pascal VOC (Everingham et al. 2010) dataset consists of 16551 images from the real world and has 20 categories. It is generally employed as the source domain. As the target domain, there are a total of three datasets, namely Clipart1K, Watercolor2K, and Comic2K (Inoue et al. 2018).

Methods	aero	bcycle	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	hrs	bike	prsn	plnt	sheep	sofa	train	tv	mAP
Source Only	27.3	60.4	17.5	16.0	14.5	43.7	32.0	10.2	38.6	15.3	24.5	16.0	18.4	49.5	30.7	30.0	2.3	23.0	35.1	29.9	26.7
WST	30.8	65.5	18.7	23.0	24.9	57.5	40.2	10.9	38.0	25.9	36.0	15.6	22.6	66.8	52.1	35.3	1.0	34.6	38.1	39.4	33.8
BSR	26.3	56.8	21.9	20.0	24.7	55.3	42.9	11.4	40.5	30.5	25.7	17.3	23.2	66.9	50.9	35.2	11.0	33.2	47.1	38.7	34.0
BSR+WST	28.0	64.5	23.9	19.0	21.9	64.3	43.5	**16.4**	42.2	25.9	30.5	7.9	25.5	67.6	54.5	36.4	10.3	31.2	57.4	43.5	35.7
SWDA	29.0	60.7	25.0	20.4	24.6	55.4	36.1	13.1	41.2	38.3	30.3	17.0	21.2	55.2	50.4	36.6	10.6	38.4	49.2	41.2	34.7
HTCN	28.7	67.7	25.3	16.1	28.7	56.0	38.9	12.5	41.0	33.0	29.6	12.9	22.9	69.0	55.9	36.1	11.8	34.1	48.8	46.8	35.8
I^3Net	30.0	67.0	32.5	21.8	29.2	62.5	41.3	11.6	37.1	39.4	27.4	19.3	25.0	67.4	55.2	**42.9**	**19.5**	36.2	50.7	39.3	37.8
DBGL	23.2	65.5	30.1	18.3	24.6	**67.6**	43.9	15.1	38.7	36.4	31.3	20.2	25.0	**74.3**	55.1	38.2	12.5	41.0	49.1	43.9	37.7
IDF	28.1	63.2	30.4	19.7	26.3	63.7	39.8	9.5	42.3	46.9	39.6	17.6	25.4	59.3	57.9	37.1	15.3	39.8	53.7	46.1	38.1
AT	30.1	**68.9**	29.8	27.8	28.8	67.1	45.1	14.3	43.3	47.2	35.9	20.3	30.1	65.1	57.3	37.7	13.3	38.7	43.2	42.1	39.3
LODS	**32.1**	65.1	**34.7**	29.9	**30.6**	63.9	44.6	16.3	46.7	46.5	43.5	19.7	28.9	58.5	64.8	40.1	14.7	36.4	39.6	41.2	39.9
CMT	31.9	66.9	33.9	**30.2**	26.3	65.2	43.6	12.6	44.5	46.3	47.9	19.3	29.9	53.1	63.3	40.5	17.1	**41.2**	**49.6**	43.9	40.4
TFD (Ours)	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**

Table 1: Results on adaptation from Pascal VOC to Clipart1K (%). All experiments are conducted based on SSD.

Methods	bike	bird	car	cat	dog	person	mAP
Source Only	77.5	46.1	44.6	30.0	26.0	58.6	47.1
WST	77.8	48.0	45.2	30.4	29.5	64.2	49.2
BSR	82.8	43.2	49.8	29.6	27.6	58.4	48.6
BSR+WST	75.6	45.8	49.3	34.1	30.3	64.1	49.9
SWDA	73.9	48.6	44.3	36.2	31.7	62.1	49.5
HTCN	78.6	47.5	45.6	35.4	31.0	62.2	50.1
I^3Net	81.1	49.3	46.2	35.0	31.9	**65.7**	51.5
DBGL	84.0	46.7	45.5	36.2	**35.7**	63.7	52.0
IDF	86.1	45.6	47.8	35.7	34.2	63.3	52.1
AT	85.8	49.6	48.3	37.8	32.5	63.3	52.9
LODS	84.1	50.5	48.7	38.1	31.4	65.2	53.0
CMT	87.1	48.7	**50.2**	37.1	31.5	66.3	53.5
TFD (Ours)	**93.0**	**52.6**	47.6	**39.2**	33.7	63.9	**55.0**

Table 2: Results on adaptation from Pascal VOC to Watercolor2K (%). All experiments are conducted based on SSD.

Methods	bike	bird	car	cat	dog	person	mAP
Source Only	43.3	9.4	23.6	9.8	10.9	34.2	21.9
WST	45.7	9.3	30.4	9.1	10.9	46.9	25.4
BSR	45.2	15.8	26.3	9.9	15.8	39.7	25.5
BSR+WST	50.6	13.6	31.0	7.5	16.4	41.4	26.8
SWDA	47.4	12.9	29.5	12.7	19.1	44.1	27.6
HTCN	50.3	15.0	27.1	9.4	18.9	46.2	27.8
I^3Net	47.5	**19.9**	33.2	11.4	19.4	49.1	30.1
DBGL	45.4	15.9	24.8	11.5	29.4	**55.1**	30.4
IDF	46.7	18.5	31.1	15.5	25.7	48.6	31.0
AT	49.4	18.3	32.1	15.2	28.7	50.3	32.3
LODS	50.6	18.6	31.4	13.9	28.9	49.8	32.2
CMT	49.8	19.2	29.8	15.2	29.1	54.1	32.9
TFD (Ours)	**53.4**	19.2	**35.0**	**16.1**	**33.2**	49.2	**34.4**

Table 3: Results on adaptation from Pascal VOC to Comic2K(%). All experiments are conducted based on SSD.

Because none of them are real-world images, there is a significant difference in data distribution from the source domain. Meanwhile, Clipart1K contains 1K images and has the same 20 categories as Pascal VOC. Both Watercolor2K and Comic2K involve 2K images and have the same 6 categories, which also exist in the source domain.

Following previous approaches (Kim et al. 2019; Chen et al. 2021b), we use the SSD (Liu et al. 2016) framework with VGG-16 architectures. VGG-16 has been pre-trained on the ImageNet dataset. The height and weight of the input image are resized to 300 pixels. The network is trained by a stochastic gradient descent optimizer with 5×10^{-4} weight decay and 0.9 momentum. The initial learning rate is 1×10^{-3}. Furthermore, the batch size is 16, consisting of 8 source images and 8 target images. All experiments are performed on one NVIDIA GeForce RTX 3090 GPU. On the target domain, we evaluate our model using mean average precision (mAP) with an IoU threshold of 0.5. We set λ_1 = 1, λ_2 = 0.5 and λ_3 = 1 in Equation (11) for all experiments.

Experimental Results

We compare the proposed TFD with several state-of-the-art DAOD methods, including BSR+WST (Kim et al. 2019), SWDA (Saito et al. 2019), HTCN (Chen et al. 2020), I^3Net (Chen et al. 2021b), DBGL (Chen et al. 2021a), IDF (Lang et al. 2022), AT (Li et al. 2022c), LODS (Li et al. 2022b) and CMT (Cao et al. 2023). The results of SWDA and HTCN are cited from (Chen et al. 2021b), and following (Chen et al. 2021b), we reproduce the IDF, AT, LODS and CMT models on our one-stage scenarios. Source Only denotes that SSD (Liu et al. 2016) is trained on the source domain and tested directly on the target domain.

Results on Clipart1K. As shown in Table 1, our proposed method achieves outstanding performance compared to other methods. TFD outperforms all compared methods in terms of mAP and advances SOTA by 0.8% (from 40.4% to 41.2%).

Results on Watercolor2K and Comic2K. Tables 2 and 3 report results for Pascal VOC to Watercolor2K and Pascal VOC to Comic2k, respectively. TFD achieves improved performance on both datasets, improving by 1.5% (53.5% to 55.0%) and 1.5% (32.9% to 34.4%), respectively.

Qualitative results. The detection results of Source Only, I^3Net and TFD on three datasets are displayed in Figure 5. The figure shows that TFD outperforms the other two detectors in terms of the ability to detect objects. Specifically, in addition to detecting objects missed by conventional detectors (e.g., (b) and (d)), the TFD also enhances the accuracy

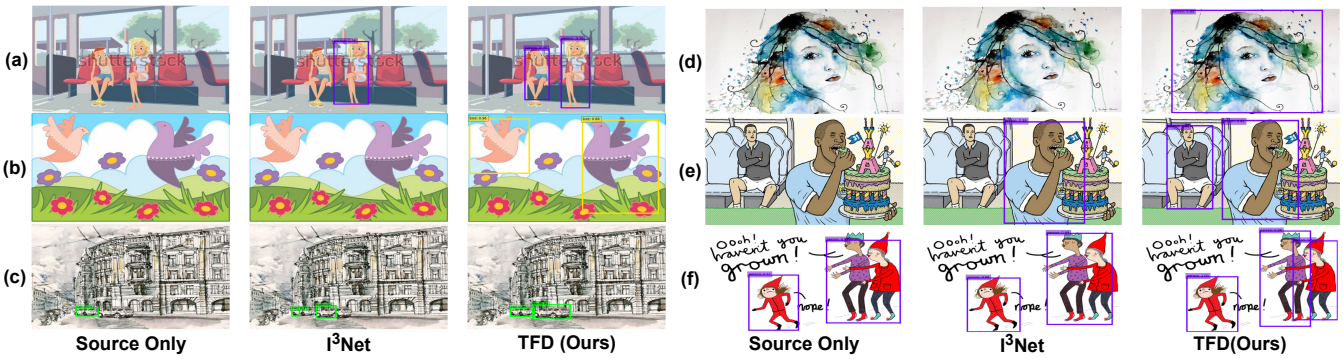

Figure 5: Detection results on Clipart1K ((a), (b)), Watercolor2K ((c), (d)) and Comic2K ((e), (f)) by Source Only (Liu et al. 2016), I^{3}Net (Chen et al. 2021b) and TFD (Ours).

Methods	Clipart1K	Watercolor2K	Comic2K
Source Only	26.7	47.1	21.9
+ SD	35.6	50.7	29.4
+ RB	37.9	52.2	32.4
+ CDM	39.0	53.6	33.1
+ CFDM$_{p1}$	39.9	54.2	33.7
+ CFDM$_{p2}$	**41.2**	**55.0**	**34.4**

Table 4: Ablation results on three datasets (%). The subscripts $p1$ and $p2$ represent part 1 and part 2 in CFDM

Methods	Clipart1K	Watercolor2K	Comic2K
(a)	40.8	54.7	34.0
(b)	41.0	54.9	**34.4**
TFD (ours)	**41.2**	**55.0**	**34.4**

Table 5: Ablation results (mAP) of RB(%). (a) use three RBs, (b) use two RBs

of bounding box predictions (e.g., (c)), leading to a notable increase in mAP.

Ablation Analysis

Effectiveness of network components. We conduct ablation experiments to further analyze the significance of proposed modules or blocks. We utilize Source Only as the benchmark. Table 4 shows the results of the ablation study. SD refers to the fact that we only introduce simple disentanglement. The experimental results clearly show that the addition of RB leads to a substantial enhancement in the performance of the detector, thereby further highlighting the significance of RB. The results also verify the significance of CDM. For example, for the Watercolor2K dataset, the performance is improved from 52.2% to 53.6%. The subscripts $p1$ and $p2$ represent part 1 and part 2 in CFDM (as shown in Figure 2), and both of them can improve the detection performance.

Influence of RB. The excellent results observed in the ablation experiment involving our RB naturally lead to the question of whether continuous RB stacking can yield even better results. Therefore, we conducted additional experiments to investigate the impact of more RB. As demonstrated in Table 5, (a) denotes the incorporation of two additional RBs, while (b) corresponds to the inclusion of one additional RB. Notably, our observations reveal that the progressive incorporation of RBs does not yield performance enhancements; rather, it results in a deterioration of performance. We attribute this observation to the inherent limitations of the ability of the network. Moreover, the continuous stacking of RBs leads to an increase in network output, with each feature inputting to RB generating two outputs. This increases model complexity and memory usage in the training stage. Based on these results, our model stands as the optimal choice.

Time step	Clipart1K	Watercolor2K	Comic2K
1	40.6	54.2	33.6
2	**41.2**	**55.0**	**34.4**
3	40.8	54.6	34.1

Table 6: The effect of time step on CDM (%).

Influence of time step. For CDM, we also conducted experiments to determine the impact of the time step of the recurrent process on the overall performance of the model, as shown in Table 6. Experiments reveal that when $T = 2$, the result is the best. The performance of the model starts to saturate when T is set to a value greater than 2. Therefore, we choose $T = 2$ in our experiments.

Conclusion

In this work, we proposed Triple Feature Disentanglement to solve the domain adaptive object detection problem based on the one-stage detectors. We suggested Multi-level Disentanglement Module (MDM) to gradually extract domain-invariant representations (DIR). Based on the MDM, a Cyclic Disentanglement Module facilitated the separation of domain-specific representations, while DIR contained fewer domain-private features. In addition, the Categorical Features Disentanglement Module was proposed to achieve category-level disentanglement and category alignment. Experimental results on three public datasets demonstrated that our method outperforms state-of-the-art ones.

Acknowledgments

This work was supported by the National Key R&D Program of China (Grant No. 2021YFA1003004), the National Natural Science Foundation of China (Grant No. 62001167, 61731009, 61961160734, 62371190).

References

Araki, R.; Onishi, T.; Hirakawa, T.; Yamashita, T.; and Fujiyoshi, H. 2020. MT-DSSD: Deconvolutional single shot detector using multi task learning for object detection, segmentation, and grasping detection. In *Proceedings of the IEEE International Conference on Robotics and Automation*, 10487–10493.

Arjovsky, M.; Chintala, S.; and Bottou, L. 2017. Wasserstein generative adversarial networks. In *Proceedings of the International conference on machine learning*, 214–223.

Cai, Q.; Pan, Y.; Ngo, C.-W.; Tian, X.; Duan, L.; and Yao, T. 2019a. Exploring object relation in mean teacher for cross-domain detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 11457–11466.

Cai, R.; Li, Z.; Wei, P.; Qiao, J.; Zhang, K.; and Hao, Z. 2019b. Learning disentangled semantic representation for domain adaptation. In *Proceedings of the International Joint Conference on Artificial Intelligence*, volume 2019, 2060.

Cao, S.; Joshi, D.; Gui, L.-Y.; and Wang, Y.-X. 2023. Contrastive Mean Teacher for Domain Adaptive Object Detectors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 23839–23848.

Chen, C.; Li, J.; Zheng, Z.; Huang, Y.; Ding, X.; and Yu, Y. 2021a. Dual Bipartite Graph Learning: A General Approach for Domain Adaptive Object Detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2703–2712.

Chen, C.; Zheng, Z.; Ding, X.; Huang, Y.; and Dou, Q. 2020. Harmonizing transferability and discriminability for adapting object detectors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8869–8878.

Chen, C.; Zheng, Z.; Huang, Y.; Ding, X.; and Yu, Y. 2021b. I3Net: Implicit instance-invariant network for adapting one-stage object detectors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 12576–12585.

Chen, Y.; Li, W.; Sakaridis, C.; Dai, D.; and Van Gool, L. 2018. Domain adaptive Faster R-CNN for object detection in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 3339–3348.

Chen, Y.; Wang, H.; Li, W.; Sakaridis, C.; Dai, D.; and Van Gool, L. 2021c. Scale-aware domain adaptive Faster R-CNN. *International Journal of Computer Vision*, 129(7): 2223–2243.

Everingham, M.; Van Gool, L.; Williams, C. K.; Winn, J.; and Zisserman, A. 2010. The pascal visual object classes (voc) challenge. *International Journal of Computer Vision*, 88(2): 303–338.

Ganin, Y.; and Lempitsky, V. 2015. Unsupervised domain adaptation by backpropagation. In *Proceedings of the International Conference on Machine Learning*, 1180–1189.

He, C.; Li, R.; Li, S.; and Zhang, L. 2022. Voxel set transformer: A set-to-set approach to 3D object detection from point clouds. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8417–8427.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2015. Spatial pyramid pooling in deep convolutional networks for visual recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(9): 1904–1916.

Inoue, N.; Furuta, R.; Yamasaki, T.; and Aizawa, K. 2018. Cross-domain weakly-supervised object detection through progressive domain adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 5001–5009.

Jiang, Z.; Zhao, L.; Li, S.; and Jia, Y. 2020. Real-time object detection method based on improved YOLOv4-tiny. *arXiv preprint arXiv:2011.04244*.

Jin, X.; Lan, C.; Zeng, W.; and Chen, Z. 2021. Style normalization and restitution for domain generalization and adaptation. *IEEE Transactions on Multimedia*, 24: 3636–3651.

Kim, J.-a.; Sung, J.-Y.; and Park, S.-h. 2020. Comparison of Faster-RCNN, YOLO, and SSD for real-time vehicle type recognition. In *Proceedings of the IEEE International Conference on Consumer Electronics-Asia*, 1–4.

Kim, S.; Choi, J.; Kim, T.; and Kim, C. 2019. Self-training and adversarial background regularization for unsupervised domain adaptive one-stage object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 6092–6101.

Lang, Q.; Zhang, L.; Shi, W.; Chen, W.; and Pu, S. 2022. Exploring implicit domain-invariant features for domain adaptive object detection. *IEEE Transactions on Circuits and Systems for Video Technology*, 33(4): 1816–1826.

Li, J.; Wang, Z.; Gao, Y.; and Hu, X. 2022a. Exploring high-quality target domain information for unsupervised domain adaptive semantic segmentation. In *Proceedings of the 30th ACM International Conference on Multimedia*, 5237–5245.

Li, S.; Ye, M.; Zhu, X.; Zhou, L.; and Xiong, L. 2022b. Source-free object detection by learning to overlook domain style. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8014–8023.

Li, Y.-J.; Dai, X.; Ma, C.-Y.; Liu, Y.-C.; Chen, K.; Wu, B.; He, Z.; Kitani, K.; and Vajda, P. 2022c. Cross-domain adaptive teacher for object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7581–7590.

Liu, D.; Zhang, C.; Song, Y.; Huang, H.; Wang, C.; Barnett, M.; and Cai, W. 2022. Decompose to adapt: Cross-domain object detection via feature disentanglement. *IEEE Transactions on Multimedia*, 25: 1333–1344.

Liu, F.; Zhang, X.; Wan, F.; Ji, X.; and Ye, Q. 2021a. Domain contrast for domain adaptive object detection. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(12): 8227–8237.

Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S.; Fu, C.-Y.; and Berg, A. C. 2016. SSD: Single shot multibox detector. In *Proceedings of the European Conference on Computer Vision*, 21–37.

Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; and Guo, B. 2021b. Swin Transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 10012–10022.

Ma, X.; Guo, J.; Sansom, A.; McGuire, M.; Kalaani, A.; Chen, Q.; Tang, S.; Yang, Q.; and Fu, S. 2021. Spatial pyramid attention for deep convolutional neural networks. *IEEE Transactions on Multimedia*, 23: 3048–3058.

Peng, X.; Huang, Z.; Sun, X.; and Saenko, K. 2019. Domain agnostic learning with disentangled representations. In *Proceedings of the International Conference on Machine Learning*, 5102–5112.

Ren, S.; He, K.; Girshick, R.; and Sun, J. 2015. Faster R-CNN: Towards real-time object detection with region proposal networks. In *Proceedings of the Advances in Neural Information Processing Systems*, volume 28.

Saito, K.; Ushiku, Y.; Harada, T.; and Saenko, K. 2019. Strong-weak distribution alignment for adaptive object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6956–6965.

Sultana, F.; Sufian, A.; and Dutta, P. 2020. A review of object detection models based on convolutional neural network. *Intelligent Computing: Image Processing Based Pplications*, 1–16.

Vs, V.; Gupta, V.; Oza, P.; Sindagi, V. A.; and Patel, V. M. 2021. MEGA-CDA: Memory guided attention for category-aware unsupervised domain adaptive object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4516–4526.

Woo, S.; Park, J.; Lee, J.-Y.; and Kweon, I. S. 2018. CBAM: Convolutional block attention module. In *Proceedings of the European Conference on Computer Vision*, 3–19.

Wu, A.; Liu, R.; Han, Y.; Zhu, L.; and Yang, Y. 2021. Vector-decomposed disentanglement for domain-invariant object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 9342–9351.

Wu, Z.; Liang, T.; Meng, M.; Liu, J.; Yu, J.; and Wu, J. 2022. Triple disentangling network for unsupervised domain adaptation. In *Proceedings of the IEEE International Conference on Multimedia and Expo*, 1–6.

Xu, S.; Zhang, H.; Xu, X.; Hu, X.; Xu, Y.; Dai, L.; Choi, K.-S.; and Heng, P.-A. 2022a. Representative Feature Alignment for Adaptive Object Detection. *IEEE Transactions on Circuits and Systems for Video Technology*, 33(2): 689–700.

Xu, Y.; Sun, Y.; Yang, Z.; Miao, J.; and Yang, Y. 2022b. H2FA R-CNN: Holistic and hierarchical feature alignment for cross-domain weakly supervised object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 14329–14339.

Yang, J.; Dvornek, N. C.; Zhang, F.; Chapiro, J.; Lin, M.; and Duncan, J. S. 2019. Unsupervised domain adaptation via disentangled representations: Application to cross-modality liver segmentation. In *Proceedings of the Medical Image Computing and Computer Assisted Intervention*, 255–263.

Yao, X.; Zhao, S.; Xu, P.; and Yang, J. 2021. Multi-source domain adaptation for object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 3273–3282.

Yin, N.; Shen, L.; Li, B.; Wang, M.; Luo, X.; Chen, C.; Luo, Z.; and Hua, X.-S. 2022. DEAL: An unsupervised domain adaptive framework for graph-level classification. In *Proceedings of the 30th ACM International Conference on Multimedia*, 3470–3479.

Yu, Y.; Zhai, Y.; and Zhang, Y. 2022. Align and adapt: a two-stage adaptation framework for unsupervised domain adaptation. In *Proceedings of the 30th ACM International Conference on Multimedia*, 4723–4732.

Zhang, H.; Luo, G.; Li, J.; and Wang, F.-Y. 2021a. C2FDA: Coarse-to-fine domain adaptation for traffic object detection. *IEEE Transactions on Intelligent Transportation Systems*, 23(8): 12633–12647.

Zhang, P.; Zhang, B.; Zhang, T.; Chen, D.; Wang, Y.; and Wen, F. 2021b. Prototypical pseudo label denoising and target structure learning for domain adaptive semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 12414–12424.

Zhao, L.; and Wang, L. 2022. Task-specific inconsistency alignment for domain adaptive object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 14217–14226.

Zhao, Q.; Lyu, S.; Liu, B.; Chen, L.; and Zhao, H. 2023. Self-training guided disentangled adaptation for cross-domain remote sensing image semantic segmentation. *arXiv preprint arXiv:2301.05526*.

Zhou, L.; Ye, M.; Li, X.; Zhu, C.; Liu, Y.; and Li, X. 2020. Disentanglement then reconstruction: Learning compact features for unsupervised domain adaptation. *arXiv preprint arXiv:2005.13947*.

Zhu, X.; Pang, J.; Yang, C.; Shi, J.; and Lin, D. 2019. Adapting object detectors via selective cross-domain alignment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 687–696.

Zhu, X.; Su, W.; Lu, L.; Li, B.; Wang, X.; and Dai, J. 2021. Deformable DETR: Deformable transformers for end-to-end object detection. In *9th International Conference on Learning Representations*.

Zhu, Y.; Sun, X.; Diao, W.; Li, H.; and Fu, K. 2022. RFA-Net: Reconstructed feature alignment network for domain adaptation object detection in remote sensing imagery. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 15: 5689–5703.