

Vector-Decomposed Disentanglement for Domain-Invariant Object Detection

Aming Wu^{1,2*} Rui Liu^{1,2*} Yahong Han^{1,2,3†} Linchao Zhu⁴ Yi Yang⁴

¹College of Intelligence and Computing, Tianjin University, Tianjin, China

²Tianjin Key Lab of Machine Learning, Tianjin University, Tianjin, China

³Peng Cheng Laboratory, Shenzhen, China

⁴ReLER Lab, AAII, University of Technology Sydney

{tjwam, ruiliu, yahong}@tju.edu.cn, {Linchao.Zhu, yi.yang}@uts.edu.au

Abstract

To improve the generalization of detectors, for domain adaptive object detection (DAOD), recent advances mainly explore aligning feature-level distributions between the source and single-target domain, which may neglect the impact of domain-specific information existing in the aligned features. Towards DAOD, it is important to extract domain-invariant object representations. To this end, in this paper, we try to disentangle domain-invariant representations from domain-specific representations. And we propose a novel disentangled method based on vector decomposition. Firstly, an extractor is devised to separate domain-invariant representations from the input, which are used for extracting object proposals. Secondly, domain-specific representations are introduced as the differences between the input and domain-invariant representations. Through the difference operation, the gap between the domain-specific and domain-invariant representations is enlarged, which promotes domain-invariant representations to contain more domain-irrelevant information. In the experiment, we separately evaluate our method on the single- and compound-target case. For the single-target case, experimental results of four domain-shift scenes show our method obtains a significant performance gain over baseline methods. Moreover, for the compound-target case (i.e., the target is a compound of two different domains without domain labels), our method outperforms baseline methods by around 4%, which demonstrates the effectiveness of our method.

1. Introduction

Though object detection has achieved many advances [27, 9, 42, 19, 26, 21], when the training and test data are from different domains, these methods usually suffer

*Equal contributions

†Corresponding author

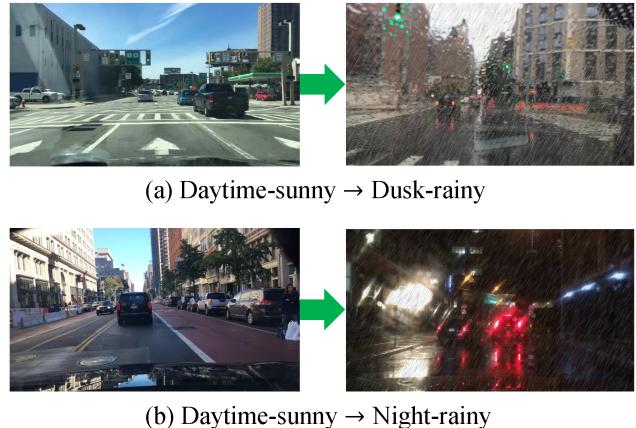
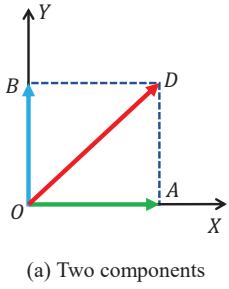


Figure 1. To further verify the proposed method, we construct two new adaptive scenes with different weather conditions.

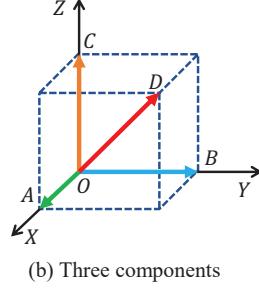
from poor generalization. To this end, the task of domain adaptive object detection (DAOD) [4] has been proposed, in which a domain gap always exists between the training/source and test/target domain, e.g., different weather conditions (as shown in Fig. 1).

To address DAOD, many methods [31, 18, 35, 46] explored to reduce the domain gap by aligning the feature-level distribution of the source and single-target domain, which may neglect the impact of the domain-specific information existing in the aligned features. Towards DAOD, it is important to obtain domain-invariant representations (DIR), which is a bridge to alleviate the domain-shift impact and can help extract domain-invariant object features.

In this paper, we focus on extracting DIR. We explore to employ disentangled representation learning (DRL) [1, 24] to disentangle DIR from domain-specific representations (DSR). As a method of feature decomposition, the purpose of DRL is to uncover a set of independent factors that give rise to the current observation [6]. And these factors should contain all the information in the observation. Inspired by the idea, we explore to utilize DRL to solve DAOD and propose a novel disentangled method to extract



(a) Two components



(b) Three components

Figure 2. Two examples of vector decomposition. (a) is the case of two components (i.e., $\overrightarrow{OD} = \overrightarrow{OA} + \overrightarrow{OB}$). (b) is the case of three components (i.e., $\overrightarrow{OD} = \overrightarrow{OA} + \overrightarrow{OB} + \overrightarrow{OC}$). Here, all these components are kept orthogonal.

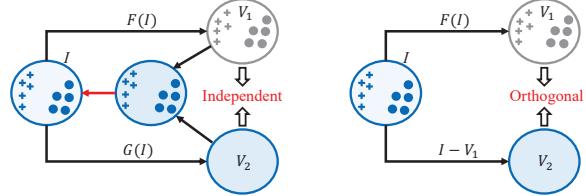
DIR. Particularly, we cast DRL into a process of vector decomposition. Vector decomposition is the general process of breaking one vector into two or more vectors that add up to the original vector, which is similar in spirit to the process of disentanglement [13]. Thus we consider employing the idea of vector decomposition to conduct disentanglement.

Concretely, given a feature map extracted by a backbone, an extractor consisting of multiple convolutional layers is devised to separate DIR from the feature map. Next, we take the difference between the feature map and DIR as DSR. Meanwhile, a domain classifier is used to help DSR contain much more domain-specific information. Besides, one key-step of disentanglement is to keep DIR and DSR independent. In this paper, we enhance independence via a constraint of vector orthogonalization between the DIR and DSR. Finally, a region proposal network (RPN) is utilized to extract object proposals from DIR. Moreover, since the proposed method is a new feature decomposition mechanism, we should design a proper optimization to obtain DIR. To this end, based on the purpose of DRL, we break DRL into two sequential training steps, i.e., the step of feature decomposition aiming at learning disentanglement, and the step of feature orthogonalization aiming at promoting DIR and DSR to be independent. The two-step optimization could promote our model learns feature decomposition, which is beneficial for extracting DIR for DAOD.

In the experiment, we first evaluate our method on the single-target case. Next, we evaluate our method on the compound-target case [23], i.e., the target is a compound of two different domains without domain labels. The significant performance gain over baselines shows the effectiveness of our disentangled method. Our code will be available at <https://github.com/AmingWu/VDD-DAOD>.

The contributions are summarized as follows:

- (1) Different from traditional disentanglement, we present a vector-decomposed disentanglement, which does not rely on the reconstruction operation to ensure the decomposed components contain all the information of input.
- (2) Based on vector-decomposed disentanglement, we



(a) Traditional Disentanglement (b) Vector-Decomposed Disentanglement

Figure 3. Comparisons between the traditional method and our vector-decomposed method. Given an input I , traditional disentanglement usually employs two extractors F and G to disentangle V_1 and V_2 . And V_1 and V_2 should be kept independent. To promote V_1 and V_2 to contain all the information of I , a reconstruction operation is usually employed. Here, the red arrow indicates the reconstruction operation. For vector-decomposed disentanglement, it only uses an extractor to decompose V_1 . The difference between I and V_1 is taken as V_2 . Meanwhile, V_1 and V_2 are kept orthogonal. Besides, vector-decomposed disentanglement does not need to utilize the reconstruction operation to promote V_1 and V_2 to contain all the information of I .

design a new framework to solve DAOD. Meanwhile, we design a two-step training strategy to optimize our model.

(3) In the experiment, our method is separately evaluated on the single- and compound-target cases. And we build two new adaptive scenes (see Fig. 1), i.e., Daytime-sunny → Dusk-rainy and Daytime-sunny → Night-rainy, to further verify our method. The significant performance gain over baselines shows the effectiveness of our method.

2. Related Work

Domain Adaptive Object Detection. Most existing methods [39, 3, 40, 32, 44, 43] employ holistic representations to align the feature- or pixel-level distributions of the source and target domain. Particularly, Chen et al. [4] proposed to align the global feature distributions to reduce the domain gap. Saito et al. [29] proposed to align the local and global feature distributions to alleviate the domain-shift impact. Besides, the work [16] utilized an encoder-decoder network to translate the style of the source domain to that of the target domain, which could be thought of as aligning the pixel-level distributions of the source and target domain. Although these methods have been demonstrated to be effective, they neglect the impact of domain-specific information existing in the aligned features, which may affect the adaptation performance. To this end, we focus on extracting domain-invariant representations for DAOD.

Disentangled Representation Learning. As an effective mechanism of feature decomposition, recently, DRL [22, 2] has been demonstrated to be effective in many tasks, e.g., image translation [17] and few-shot learning [28]. Particularly, the work [17] employs DRL to decompose DSR to make diverse image style translation. Peng et al. [25] utilize DRL to disentangle three different factors to make domain

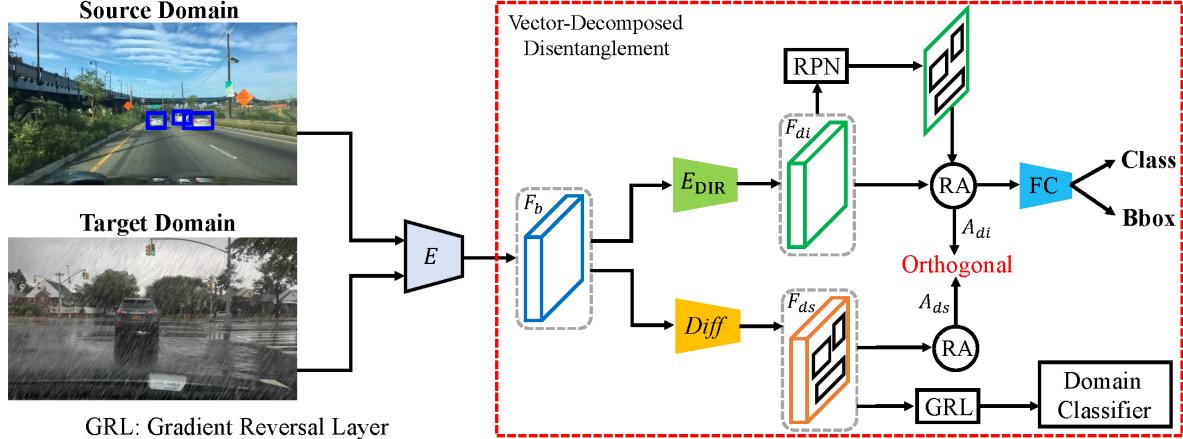


Figure 4. Illustration of vector-decomposed disentanglement: a plug component for domain adaptive Faster R-CNN series [29, 39]. ‘RA’ and ‘Diff’ separately indicate Roi-Alignment and the difference decomposition. ‘GRL’ denotes Gradient Reversal Layer [4, 29]. We first design an extractor E_{DIR} to decompose DIR (F_{di}) from F_b . Then, based on F_{di} , RPN is employed to extract domain-invariant proposals.

adaptive classification. However, since this work only considers holistic image-level representations for classification, it could not be applied directly to object detection.

In this paper, we consider DRL from the perspective of vector decomposition. Particularly, our method only requires to devise an extractor to decompose DIR. And DSR could be obtained from the difference between the input and DIR. Experimental results on single- and compound-target DAOD demonstrate the effectiveness of our method.

3. Vector-Decomposed Disentanglement

As discussed in the section of Introduction, the purpose of vector decomposition is to break one vector into two or more components that add up to the original vector. In general, each vector can be taken as the sum of two or more other vectors. Fig. 2 shows two decomposed examples, i.e., $\overrightarrow{OD} = \overrightarrow{OA} + \overrightarrow{OB}$ and $\overrightarrow{OD} = \overrightarrow{OA} + \overrightarrow{OB} + \overrightarrow{OC}$.

Obviously, vector decomposition is similar in spirit to disentanglement. And the decomposition idea is also applied to high-dimensional space. Therefore, we consider employing vector decomposition to obtain disentangled representations. Concretely, for the case of two components (Fig. 3(b)), give an input representation I , we design an extractor F to decompose the first component V_1 from I . Then, we take the difference between I and V_1 as the second component V_2 . Here, we name the process extracting V_2 as difference decomposition.

$$V_1 = F(I), V_2 = I - V_1, V_1 \perp V_2, \quad (1)$$

where \perp indicates two components are orthogonal.

Compared with the traditional disentanglement (Fig. 3(a)), vector decomposition only takes the difference between the original input and decomposed components as the last component, which reduces parameters and computational costs. Moreover, the difference decomposition of

obtaining the last component could make all the components contain all the information of the input, which does not rely on the reconstruction operation. In the following, we will introduce the details of vector-decomposed disentanglement for domain adaptive object detection.

4. Domain-Invariant Object Detection

For DAOD, we could access image x^s with labels y^s and bounding boxes b^s , which are from the source domain. And we could also access image x^t that is from the target domain. The goal is to obtain the results of the target domain.

4.1. The Network of Disentanglement

The right part of Fig. 4 illustrates the details of vector-decomposed disentanglement, which is plugged into the domain adaptive Faster R-CNN series [29, 39, 27]. Concretely, given an image x^s and x^t , we first obtain a feature map F_b that is the output of a feature extractor E . Next, we define an extractor E_{DIR} to decompose domain-invariant feature F_{di} from F_b . And the difference between F_b and F_{di} is taken as the domain-specific feature F_{ds} .

$$F_{di} = E_{DIR}(F_b), F_{ds} = F_b - F_{di}. \quad (2)$$

Here, E_{DIR} indicates the DIR extractor. The size of F_{di} and F_{ds} is set to the same as that of F_b . Next, a Region Proposal Network (RPN) is performed on F_{di} to extract a set of domain-invariant proposals. Finally, for an image from the source domain, the detection loss is defined as follows:

$$\mathcal{L}_{det} = \mathcal{L}_{loc} + \mathcal{L}_{cls} + \mathcal{L}_{rpn}, \quad (3)$$

where \mathcal{L}_{loc} and \mathcal{L}_{cls} separately indicate the bounding-box regression loss and classification loss. \mathcal{L}_{rpn} is the loss of RPN to distinguish foreground from background and to refine bounding-box anchors.

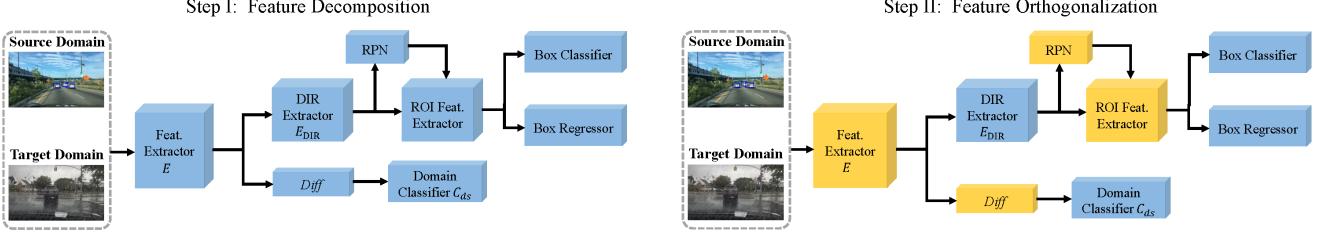


Figure 5. Illustration of our two-step optimization process. In the first training step (i.e., feature decomposition), the entire object detector (all the blue blocks) are jointly trained on the source and target images. In the second step (i.e., feature orthogonalization), the parameters in the yellow blocks are fixed. We only update the parameters in the blue blocks.

4.2. Training with the Two-step Optimization

The goal of our method (see Eq. (1)) is to decompose a set of orthogonal components. To enhance the disentangled ability, we break vector decomposition into two sequential steps. Specifically, we first promote models to be capable of decomposing components. Then, a constraint is imposed to promote these components to be orthogonal.

The step of feature decomposition. The step is to promote our model to decompose input features into two different components. Concretely, based on F_{di} , we first employ RPN to extract object proposals. Then, for a source image, the processes of detection loss are shown in Eq. (3).

Next, to promote the difference result F_{ds} to contain much more domain-specific information, we utilize the adversarial training mechanism [8] and design a network C_{ds} to perform domain classification. And the domain label D is set to 0 for the source domain and 1 for the target domain. Finally, the loss of the first step is shown as follows:

$$\begin{aligned}\mathcal{L}_{src}^1 &= \mathcal{L}_{det} + \mathcal{L}_{dom}(C_{ds}(F_{ds})), \\ \mathcal{L}_{tgt}^1 &= \mathcal{L}_{dom}(C_{ds}(F_{ds})),\end{aligned}\quad (4)$$

where \mathcal{L}_{src}^1 and \mathcal{L}_{tgt}^1 are the objective functions of the source and target domain, respectively. \mathcal{L}_{dom} is the domain classification loss, i.e., $\mathcal{L}_{dom} = -[D \log \hat{D} + (1-D) \log(1-\hat{D})]$ and $\hat{D} = C_{ds}(F_{ds})$. Finally, we take the sum of \mathcal{L}_{src}^1 and \mathcal{L}_{tgt}^1 to optimize the entire model.

The step of feature orthogonalization. In this step, we first fix the feature extractor E . Then, we use the extractor E_{DIR} to obtain F_{di} (Eq. (2)). Next, RPN is performed on F_{di} to extract a set of object proposals.

The key idea of disentanglement [6] is to keep the disentangled components independent. Here, based on the theory of vector decomposition, we try to promote the decomposed components are orthogonal, which is equivalent to the independent operation. Thus we impose an **orthogonal loss** \mathcal{L}_\perp on the DIR and DSR. Concretely, based on object proposals, we first obtain the Roi-Alignment result $A_{di} \in \mathbb{R}^{n \times c \times h \times w}$ of F_{di} and $A_{ds} \in \mathbb{R}^{n \times c \times h \times w}$ of F_{ds} , where n , c , h , and w indicate the number of proposals, the number of channels, the height and width, respectively. The process of orthogo-

nal loss is shown as follows:

$$\begin{aligned}M &= (\|P_{di}\|_2^2) \odot (\|P_{ds}\|_2^2), \\ \mathcal{L}_\perp &= \frac{1}{n} \sum_{i=1}^n \left| \sum_{j=1}^c M[i, j] \right|,\end{aligned}\quad (5)$$

where $P_{di} \in \mathbb{R}^{n \times c}$ and $P_{ds} \in \mathbb{R}^{n \times c}$ are the results of global average pooling. $\|\cdot\|_2^2$, $|\cdot|$, and \odot separately indicate L2-norm, the absolute value operation, and element-wise product. $M[i, j]$ indicates the value of $M \in \mathbb{R}^{n \times c}$ at the position (i, j) . Besides, it is worth noting that we use the alignment results instead of the overall feature map to compute the orthogonal loss, which could not only reduce computational costs but also promote our model to focus on object regions.

By minimizing the orthogonal loss, we could promote F_{di} and F_{ds} are independent. Since F_{ds} contains more domain-specific information, this loss can promote F_{di} to contain much more domain-invariant information. Finally, the loss of the second step is defined as follows:

$$\begin{aligned}\mathcal{L}_{src}^2 &= \mathcal{L}_{det} + \mathcal{L}_{dom}(C_{ds}(F_{ds})) + \mathcal{L}_\perp, \\ \mathcal{L}_{tgt}^2 &= \mathcal{L}_{dom}(C_{ds}(F_{ds})) + \mathcal{L}_\perp,\end{aligned}\quad (6)$$

where \mathcal{L}_{det} is the detection loss based on A_{di} . The sum of \mathcal{L}_{src}^2 and \mathcal{L}_{tgt}^2 is used to optimize certain components of the model. The processes are shown in the right part of Fig. 5. After the second training step, the decomposed DIR and DSR will be kept independent, which enhances the disentangled ability of our model.

In this paper, our model is trained in an end-to-end way. The training details are shown in Algorithm 1. Besides, for the second training step, the parameters that do not appear in the step are considered to be fixed.

4.3. Discussion about Learning DIR

For our method, we have two operations to promote to learn domain-invariant features. Firstly, the difference decomposition makes F_{di} contain much less domain-relevant information. Secondly, the orthogonal loss can further promote F_{di} to contain much more domain-irrelevant information. And we consider domain-irrelevant information con-

Algorithm 1 Two-step optimization for DAOD

Require:

source images $\{x^s, y^s, b^s\}$; target images $\{x^t\}$; feature extractor E ; DIR extractor E_{DIR} ; domain classifier C_{ds} .

Ensure:

feature extractor \hat{E} , DIR extractor \hat{E}_{DIR} .

- 1: **while** not converged **do**
 - 2: Sample a mini-batch from $\{x^s, y^s, b^s\}$ and $\{x^t\}$;
 - 3: **Feature Decomposition:**
 - 4: Compute $\mathcal{L}_1 = \mathcal{L}_{src}^1 + \mathcal{L}_{tgt}^1$ (Eq. (4));
 - 5: Update E , E_{DIR} , and C_{ds} by \mathcal{L}_1 ;
 - 6: Update RPN module, Classifier, and Regressor by \mathcal{L}_1 ;
 - 7: **Feature Orthogonalization:**
 - 8: Compute $\mathcal{L}_2 = \mathcal{L}_{src}^2 + \mathcal{L}_{tgt}^2$ (Eq. (6));
 - 9: Update E_{DIR} , C_{ds} by \mathcal{L}_2 ;
 - 10: Update Classifier and Regressor by \mathcal{L}_2 ;
 - 11: **end while**
 - 12: **return** $\hat{E} = E$; $\hat{E}_{DIR} = E_{DIR}$.
-

tains domain-invariant information. Thus, these two operations promote F_{di} contains much more domain-invariant information, which reduces the domain-shift impact.

5. Experiment

In the experiment, we separately evaluate our approach on single- and compound-target DAOD. For the single-target case, our method is evaluated on four domain-shift scenes, i.e., Cityscapes [5] \rightarrow FoggyCityscapes [30], PASCAL [7] \rightarrow Watercolor [14], Daytime-sunny \rightarrow Dusk-rainy, and Daytime-sunny \rightarrow Night-rainy. For the compound-target case [23], we take Daytime-sunny as the source domain and the compound of Dusk-rainy and Night-rainy as the target domain, whose goal is to adapt a model from labeled source domain to unlabeled compound target domain. All the experiments are trained in an end-to-end way.

Datasets. Cityscapes is a dataset about city street scene. It contains 2,975 images for training and 500 images for validation. FoggyCityscapes is rendered based on Cityscapes. And it shows street scene under foggy weather. We follow the setting of the work [29] and evaluate our method on the validation set. For PASCAL \rightarrow Watercolor, we utilize Pascal VOC dataset as the source domain. It contains 20 classes of images and bounding box annotations. Following the setting of the work [29], we employ Pascal VOC 2007 and 2012 training and validation splits for training, which results in about 15K images. Watercolor contains 2K images with 6 categories. The splits of the training and test set are the same as the work [29].

The Berkeley Deep Drive 100k (BDD-100k) dataset [41] consists of 100,000 driving videos. Based on this dataset, we build two new adaptive scenes. As shown in Fig. 1, for Daytime-sunny \rightarrow Dusk-rainy, we select 27,708 daytime-sunny images as the source domain and 3,501 dusk-rainy

Method	prsn	rider	car	truck	bus	train	meycl	bcycl	mAP
Source Only	24.7	31.9	33.1	11.0	26.4	9.2	18.0	27.9	22.8
DAF [4]	25.0	31.0	40.5	22.1	35.3	20.2	20.0	27.1	27.6
DT [14]	25.4	39.3	42.4	24.9	40.4	23.1	25.9	30.4	31.5
SC-DA [45]	33.5	38.0	48.5	26.5	39.0	23.3	28.0	33.6	33.8
DMRL [16]	30.8	40.5	44.3	27.2	38.4	34.5	28.4	32.2	34.6
MLDA [38]	33.2	44.2	44.8	28.2	41.8	28.7	30.5	36.5	36.0
FSDA [36]	29.1	39.7	42.9	20.8	37.4	24.1	26.5	29.9	31.3
MAF [11]	28.2	39.5	43.9	23.8	39.9	33.3	29.2	33.9	34.0
CT [43]	32.7	44.4	50.1	21.7	45.6	25.4	30.1	36.8	35.9
CDN [34]	35.8	45.7	50.9	30.1	42.5	29.8	30.8	36.5	36.6
SCL [32]	31.6	44.0	44.8	30.4	41.8	40.7	33.6	36.2	37.9
ATF [12]	34.6	47.0	50.0	23.7	43.3	38.7	33.4	38.8	38.7
MCAR [44]	32.0	42.1	43.9	31.3	44.1	43.4	37.4	36.6	38.8
HTCN [3]	33.2	47.5	47.9	31.6	47.4	40.9	32.3	37.1	39.8
SW [29]	29.9	42.3	43.5	24.5	36.2	32.6	30.0	35.3	34.3
SW-VDD (ours)	32.1	42.8	49.4	29.0	49.0	33.9	29.9	37.1	37.9
ICCR [39]	32.9	43.8	49.2	27.2	45.1	36.4	30.3	34.6	37.4
ICCR-VDD (ours)	33.4	44.0	51.7	33.9	52.0	34.7	34.2	36.8	40.0

Table 1. Results (%) on adaptation from Cityscapes to FoggyCityscapes. ‘prsn’, ‘meycl’, and ‘bcycl’ separately denote ‘person’, ‘motorcycle’, and ‘bicycle’ category. ‘VDD’ indicates vector-decomposed disentanglement.

images as the target domain. For Daytime-sunny \rightarrow Night-rainy, we select 27,708 daytime-sunny images as the source domain and 2,494 night-rainy images as the target domain. Besides, for the compound-target case, we select 27,708 daytime-sunny images as the source domain and 5,995 images consisting of dusk-rainy and night-rainy as the compound target domain. Meanwhile, we render these rainy images to enlarge the gap between the source and target domain. The number of annotation boxes is around 455,000. We evaluate the performance on the target domain. Besides, the BDD-100k dataset includes ten categories. Here, we choose seven commonly used categories, which do not include the category of light, sign, and train.

Implementation Details. We employ three convolutional layers as the domain-invariant feature extractor E_{DIR} . And we separately design a network with three fully connected layers as the domain classifiers. Finally, during training, we first train our model with learning rate 0.001 for 50K iterations, then with the learning rate 0.0001 for 30K more iterations. In the test, we utilize mean average precisions (mAP) as the evaluation metric. More details can be seen in the supplementary material.

5.1. Result Analysis of Single-target DAOD

Results on FoggyCityscapes. Table 1 shows the results of FoggyCityscapes. Here, VGG16 [33] is taken as the backbone. Through plugging our disentanglement into domain adaptive Faster R-CNN methods, the performance can be improved significantly. Particularly, for SW [29] and ICCR [39], our method separately improves the performance by 3.6% and 2.6%. This demonstrates decomposing domain-invariant features is helpful for alleviating the domain-shift impact on object detection.

The first row of Fig. 6 shows one detection example from

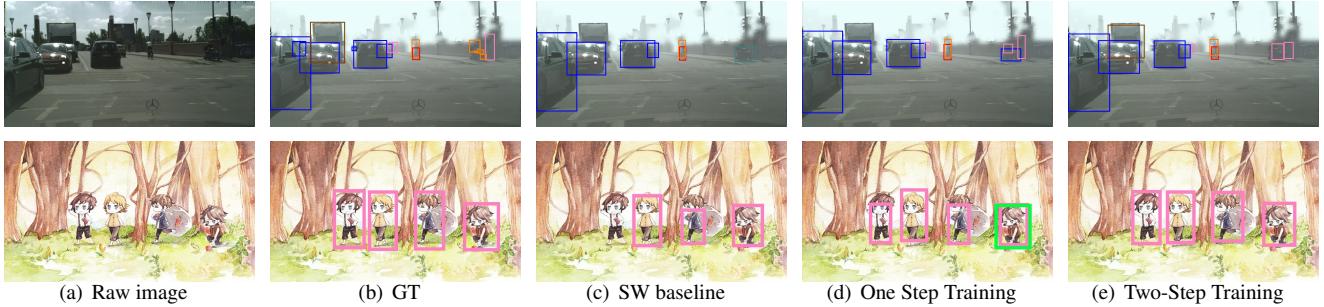


Figure 6. Detection results on the FoggyCityscapes and Watercolor scene. ‘GT’ indicates the groundtruth result. ‘One-Step Training’ denotes we integrate all loss functions and use one optimization step to train SW-VDD. We can see that using two training steps could localize and recognize objects existing in the two images accurately, e.g., the **truck**, **car**, **person**, **bicycle**, and **bird**.

Method	bike	bird	car	cat	dog	person	mAP
Source Only	68.8	46.8	37.2	32.7	21.3	60.7	44.6
BDC-Faster [29]	68.6	48.3	47.2	26.5	21.7	60.5	45.5
DAF [4]	75.2	40.6	48.0	31.5	20.6	60.0	46.0
WST-BSR [15]	75.6	45.8	49.3	34.1	30.3	64.1	49.9
MAF [11]	73.4	55.7	46.4	36.8	28.9	60.8	50.3
DC [20]	76.7	53.2	45.3	41.6	35.5	70.0	53.7
ATF [12]	78.8	59.9	47.9	41.0	34.8	66.9	54.9
SCL [32]	82.2	55.1	51.8	39.6	38.4	64.0	55.2
MCAR [44]	87.9	52.1	51.8	41.6	33.8	68.8	56.0
SW [29]	82.3	55.9	46.5	32.7	35.5	66.7	53.3
SW-VDD (ours)	90.0	56.6	49.2	39.5	38.8	65.3	56.6

Table 2. Results (%) on adaptation from Pascal to Watercolor.

the FoggyCityscapes dataset. Here, we take SW [29] as an example. We can see that compared with SW, our method localizes and recognizes objects existing in the foggy image accurately. This further shows our method is effective.

Results on Watercolor. Table 2 shows the Watercolor results. Here, we use ResNet101 [10] as the backbone. We can see plugging vector-decomposed disentanglement into SW [29] improves its performance significantly. Besides, MCAR [44] exploits multi-label object recognition as a dual auxiliary task to improve the alignment. We can see that our method outperforms MCAR by 0.6%. These all demonstrate our method is effective. And our method could alleviate the impact of the watercolor style.

The second row of Fig. 6 shows one watercolor example. We can see compared with SW, our method could localize and recognize objects accurately. These further demonstrate employing vector-decomposed disentanglement could indeed alleviate the domain-shift impact.

Results on Dusk-rainy. Table 3 shows the results of Daytime-sunny → Dusk-rainy. ResNet101 [10] is taken as the backbone. We can see that for this scene, the adaptation performance of state-of-the-art methods, e.g., CT [43] and HTCN [3], is weak. Besides, we can also see that plugging the disentanglement into SW [29] and ICCR [39] improves their performance significantly. The performance is separately improved by 5.4% and 2.9%. This further demon-

Method	bus	bike	car	motor	person	rider	truck	mAP
Source Only	38.6	21.5	51.7	12.0	19.7	13.6	40.9	28.3
CT [43]	35.5	20.3	50.9	7.9	21.6	16.1	34.4	26.7
SCL [32]	34.8	19.2	50.8	13.2	25.9	18.0	38.1	28.6
HTCN [3]	35.9	21.1	51.1	13.7	24.0	16.6	39.0	28.8
DAF [4]	43.6	27.5	52.3	16.1	28.5	21.7	44.8	33.5
SW [29]	40.0	22.8	51.4	15.4	26.3	20.3	44.2	31.5
SW-VDD	46.1	31.1	54.4	25.3	31.0	22.4	47.6	36.9
ICCR [39]	43.8	28.5	52.4	22.7	29.2	21.9	45.6	34.9
ICCR-VDD	47.9	33.2	55.1	26.1	30.5	23.8	48.1	37.8

Table 3. Results (%) on adaptation from Daytime-sunny to Dusk-rainy. Here, we directly run the released codes of the compared methods to obtain the results.

strates vector-decomposed disentanglement is capable of disentangling domain-invariant features, which is helpful for alleviating the domain-shift impact on object detection.

The first row of Fig. 7 shows three detection examples of the dusk-rainy scene. We can see that this is a challenging adaptation scene. The images are very obscure. Our method localizes and recognizes objects existing in these images accurately, which further demonstrates the effectiveness of vector-decomposed disentanglement.

Results on Night-rainy. Table 4 shows the results of Daytime-sunny → Night-rainy. ResNet101 [10] is taken as the backbone. We can see that for this scene, the performance of many adaptation methods [3, 32, 43] is weak. For example, the mAP value of HTCN and CT is lower than 20%. Plugging the disentanglement into SW [29] and ICCR [39] improves their performance significantly. The performance is improved by 5.7% and 3.1%. Particularly, for each object category, our method outperforms SW [29] and ICCR [39]. This further demonstrates the effectiveness of vector-decomposed disentanglement.

The second row of Fig. 7 shows three detection examples of the night-rainy scene. We can see for this scene, the brightness of images is very low. Meanwhile, the rainy images are very obscure. Our method localizes and recognizes objects existing in the night-rainy images accurately. This demonstrates extracting domain-invariant representations is helpful for alleviating the domain-shift impact. Our method



Figure 7. The first and second row separately show the detection results on the “Daytime-sunny → Dusk-rainy” and “Daytime-sunny → Night-rainy”. We can see our method detects objects existing in these images, which shows the effectiveness of our method.

Method	bus	bike	car	motor	person	rider	truck	mAP
Source Only	23.4	13.3	31.8	1.5	10.2	10.9	23.2	16.3
CT [43]	22.4	9.7	27.4	0.6	9.3	9.3	13.4	13.1
SCL [32]	20.0	9.2	33.2	0.3	11.9	10.6	26.4	15.9
HTCN [3]	22.8	9.4	30.7	0.7	11.9	4.8	22.0	14.6
DAF [4]	23.8	12.0	37.7	0.2	14.9	4.0	29.0	17.4
SW [29]	24.7	10.0	33.7	0.6	13.5	10.4	29.1	17.4
SW-VDD	31.7	15.3	38.0	11.1	18.2	16.7	30.8	23.1
ICCR [39]	32.5	12.1	36.2	1.3	16.1	17.0	29.3	20.6
ICCR-VDD	34.8	15.6	38.6	10.5	18.7	17.3	30.6	23.7

Table 4. Results (%) on Daytime-sunny → Night-rainy.

could extract domain-invariant representations effectively.

5.2. Ablation Analysis

Based on the single-target case, we plug our method into SW [29] to make an ablation analysis. Table 5 shows the results. We can see that for our model, employing two training steps is effective. Particularly, two-step training outperforms one-step training by 3.4% and 2.1%. This shows our optimization mechanism promotes the model to extract domain-invariant representations, which is beneficial for DAOD. In Fig. 6(d), we show two examples based on one training step. We can see using two training steps could detect objects existing in the two images accurately. Moreover, we can also see that the orthogonal loss could improve the performance significantly. This shows the orthogonal loss is indeed helpful for promoting DIR and DSR to be independent, which improves the disentangled ability.

Compared with traditional disentanglement. To further demonstrate the effectiveness of our method, we replace our method with the traditional disentanglement [25, 37]. Other components are kept unchanged. We employ the same training steps to optimize the model. Based on FoggyCityscapes and Watercolor dataset, the adaptation performance of the traditional disentanglement is 34.1% and

Method	One-step	Two-step	OL	C → F	V → W
SW-VDD	✓			33.2%	52.7%
SW-VDD	✓		✓	34.5%	54.5%
SW-VDD		✓		36.5%	54.9%
SW-VDD		✓	✓	37.9%	56.6%

Table 5. Ablation analysis of our method. Here, we use mAP as the metric. ‘One-step’ and ‘Two-step’ indicate we use one training step and two training steps to optimize our model, respectively. ‘OL’ denotes the orthogonal loss. ‘C → F’ denotes the adaptation from Cityscapes to FoggyCityscapes and employs VGG16 as the backbone. ‘V → W’ denotes the adaptation from VOC to Watercolor and utilizes ResNet101 as the backbone.

54.6%, which is weaker than our method. Besides, since our method does not include the reconstruction stage, our method owns much fewer parameters and computational costs. These all demonstrate the performance of our method outperforms the traditional disentangled method. Meanwhile, this also shows that our vector-decomposed disentanglement could extract domain-invariant features effectively, which improves the detection performance.

Visualization analysis. In Fig. 8, we compare DIR extracted by our disentangled method and traditional disentanglement. We find that compared with traditional disentanglement, the DIR extracted by our vector-decomposed disentanglement contains much less domain-specific information. Particularly, for these examples, we can see that the DIR extracted by the traditional disentanglement contains much more domain-specific information, e.g., the TD-DIR (Fig. 8(f)) of the bird image, which leads to the incorrect detections. This further demonstrates the effectiveness of our vector-decomposed disentanglement.

5.3. Result Analysis of Compound-target DAOD

For compound-target DAOD, we use the same optimization method as that of the single-target case. ResNet101

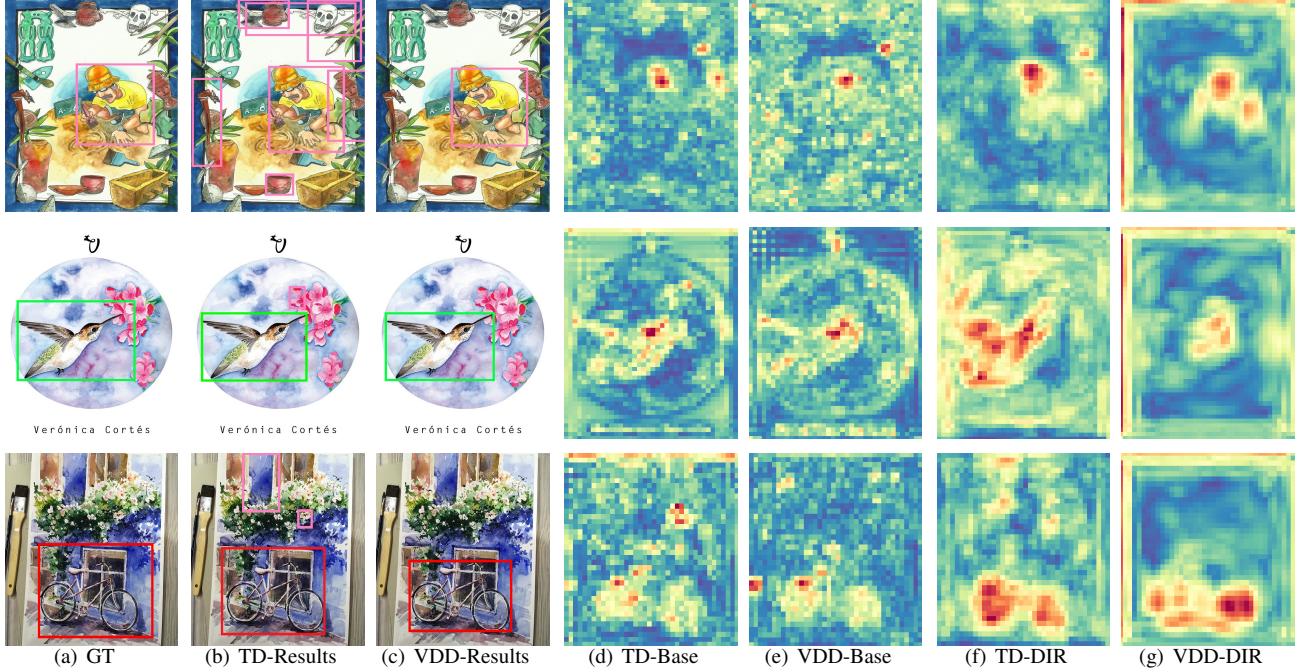


Figure 8. Comparisons of feature maps extracted by our vector-decomposed disentanglement (VDD) and traditional disentanglement (TD). Here, ‘TD-Results’ indicates detection results of TD. ‘TD-Base’ and ‘VDD-Base’ separately indicate the feature map used for disentanglement. ‘TD-DIR’ and ‘VDD-DIR’ separately indicate the DIR extracted by TD and VDD. These examples are from the ‘Pascal VOC → Watercolor’ scene. For each feature map, the channels corresponding to the maximum value are selected for visualization.

Method	bus	bike	car	motor	person	rider	truck	mAP
Source Only	35.1	19.3	44.0	8.8	17.5	12.8	37.7	25.0
DAF [4]	35.9	18.3	44.2	10.1	22.0	17.9	39.9	26.9
CT [43]	31.3	15.4	41.7	8.4	19.1	15.3	32.3	23.4
SCL [32]	32.7	19.7	44.9	10.5	22.9	18.5	38.3	26.8
SW [29]	36.9	20.7	45.1	6.6	23.1	16.9	41.5	27.3
ICCR [39]	38.8	20.4	44.6	11.7	24.7	15.4	41.6	28.2
SW-VDD	41.8	26.8	48.6	17.9	27.0	22.2	44.1	32.6

Table 6. Results (%) on the compound target domain.

Method	bus	bike	car	motor	person	rider	truck	mAP
Source Only	38.6	21.5	51.7	12.0	19.7	13.6	40.9	28.3
DAF [4]	39.5	21.0	51.6	12.6	24.8	20.5	42.7	30.4
CT [43]	34.9	17.6	49.8	11.6	21.9	17.9	35.6	27.0
SCL [32]	35.7	22.3	50.7	14.8	25.3	19.9	40.1	29.8
SW [29]	39.2	24.6	49.6	9.2	25.5	19.3	43.7	30.1
ICCR [39]	42.0	21.9	51.5	16.5	27.2	16.8	44.1	31.4
SW-VDD	43.7	30.3	52.7	22.3	29.7	24.8	46.4	35.7

Table 7. Results (%) on the dusk-rainy scene. Here, the trained model is directly evaluated on the dusk-rainy scene.

is the backbone. Table 6, 7, and 8 show the compared results. Here, the model trained on the compound-target DAOD is separately evaluated on the compound-target, dusk-rainy, and night-rainy domain. Compared with SW [29], plugging our disentanglement into SW improves its performance by 5.3%, 5.6%, and 5.0%. Meanwhile, we can see that the performance of each category outperforms all compared methods significantly. This shows for single- and compound-target DAOD, extracting DIR is an efficient way.

Method	bus	bike	car	motor	person	rider	truck	mAP
Source Only	23.4	13.3	31.8	1.5	10.2	10.9	23.2	16.3
DAF [4]	24.2	11.0	32.4	4.6	12.7	11.9	27.7	17.8
CT [43]	19.5	9.7	29.0	1.1	9.9	9.1	17.6	13.7
SCL [32]	22.9	12.8	35.8	0.9	14.8	15.0	30.2	18.9
SW [29]	29.6	10.4	37.9	0.7	15.0	11.1	31.6	19.5
ICCR [39]	28.4	16.5	33.6	0.9	16.4	12.2	30.3	19.7
SW-VDD	35.7	17.4	42.2	7.9	18.1	16.0	33.9	24.5

Table 8. Results (%) on the night-rainy scene. Here, the trained model is directly evaluated on the night-rainy scene.

Meanwhile, the performance gain further demonstrates our method is capable of extracting DIR effectively.

6. Conclusion

In this paper, we propose vector-decomposed disentanglement for DAOD. We only defines an extractor to extract domain-invariant representations. Meanwhile, we do not use reconstruction to ensure the disentangled components contain all the information in the input. In the experiment, our method is separately evaluated on the single- and compound-target case. The performance gain over baselines shows the effectiveness of our method.

Acknowledgement

This work is supported by the NSFC (under Grant 61876130, 61932009).

References

- [1] Yoshua Bengio, Aaron Courville, and Pascal Vincent. Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence*, 35(8):1798–1828, 2013.
- [2] Ruichu Cai, Zijian Li, Pengfei Wei, Jie Qiao, Kun Zhang, and Zhifeng Hao. Learning disentangled semantic representation for domain adaptation. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence*, pages 2060–2066, 2019.
- [3] Chaoqi Chen, Zebiao Zheng, Xinghao Ding, Yue Huang, and Qi Dou. Harmonizing transferability and discriminability for adapting object detectors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8869–8878, 2020.
- [4] Yuhua Chen, Wen Li, Christos Sakaridis, Dengxin Dai, and Luc Van Gool. Domain adaptive faster r-cnn for object detection in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3339–3348, 2018.
- [5] 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 *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3213–3223, 2016.
- [6] Kien Do and Truyen Tran. Theory and evaluation metrics for learning disentangled representations. *arXiv preprint arXiv:1908.09961*, 2019.
- [7] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88(2):303–338, 2010.
- [8] Yaroslav Ganin and Victor Lempitsky. Unsupervised domain adaptation by backpropagation. *arXiv preprint arXiv:1409.7495*, 2014.
- [9] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017.
- [10] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.
- [11] Zhenwei He and Lei Zhang. Multi-adversarial faster-rcnn for unrestricted object detection. *arXiv preprint arXiv:1907.10343*, 2019.
- [12] Zhenwei He and Lei Zhang. Domain adaptive object detection via asymmetric tri-way faster-rcnn. *European Conference on Computer Vision*, 2020.
- [13] Irina Higgins, David Amos, David Pfau, Sébastien Racanière, Loïc Matthey, Danilo Rezende, and Alexander Lerchner. Towards a definition of disentangled representations. *arXiv preprint arXiv:1812.02230*, 2018.
- [14] Naoto Inoue, Ryosuke Furuta, Toshihiko Yamasaki, and Kiyoharu Aizawa. Cross-domain weakly-supervised object detection through progressive domain adaptation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5001–5009, 2018.
- [15] Seunghyeon Kim, Jaehoon Choi, Taekyung Kim, and Changick Kim. 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*, pages 6092–6101, 2019.
- [16] Taekyung Kim, Minki Jeong, Seunghyeon Kim, Seokeon Choi, and Changick Kim. Diversify and match: A domain adaptive representation learning paradigm for object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 12456–12465, 2019.
- [17] Hsin-Ying Lee, Hung-Yu Tseng, Jia-Bin Huang, Maneesh Singh, and Ming-Hsuan Yang. Diverse image-to-image translation via disentangled representations. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 35–51, 2018.
- [18] Shuang Li, Chi Harold Liu, Xie Binhi, Limin Su, Zhengming Ding, and Gao Huang. Joint adversarial domain adaptation. In *Proceedings of the 27th ACM International Conference on Multimedia*, pages 729–737, 2019.
- [19] Xin Li, Fan Yang, Hong Cheng, Junyu Chen, Yuxiao Guo, and Leiting Chen. Multi-scale cascade network for salient object detection. In *Proceedings of the 25th ACM International Conference on Multimedia*, page 439447, 2017.
- [20] Feng Liu, Xiaoxong Zhang, Fang Wan, Xiangyang Ji, and Qixiang Ye. Domain contrast for domain adaptive object detection. *arXiv preprint arXiv:2006.14863*, 2020.
- [21] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. Ssd: Single shot multibox detector. In *European conference on computer vision*, pages 21–37. Springer, 2016.
- [22] Yen-Cheng Liu, Yu-Ying Yeh, Tzu-Chien Fu, Sheng-De Wang, Wei-Chen Chiu, and Yu-Chiang Frank Wang. Detach and adapt: Learning cross-domain disentangled deep representation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8867–8876, 2018.
- [23] Ziwei Liu, Zhongqi Miao, Xingang Pan, Xiaohang Zhan, Dahua Lin, Stella X Yu, and Boqing Gong. Open compound domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12406–12415, 2020.
- [24] Francesco Locatello, Stefan Bauer, Mario Lucic, Sylvain Gelly, Bernhard Schölkopf, and Olivier Bachem. Challenging common assumptions in the unsupervised learning of disentangled representations. 2019.
- [25] Xingchao Peng, Zijun Huang, Ximeng Sun, and Kate Saenko. Domain agnostic learning with disentangled representations. *ICML*, 2019.
- [26] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 779–788, 2016.
- [27] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, pages 91–99, 2015.

- [28] Karl Ridgeway and Michael C Mozer. Learning deep disentangled embeddings with the f-statistic loss. In *Advances in Neural Information Processing Systems*, pages 185–194, 2018.
- [29] Kuniaki Saito, Yoshitaka Ushiku, Tatsuya Harada, and Kate Saenko. Strong-weak distribution alignment for adaptive object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6956–6965, 2019.
- [30] Christos Sakaridis, Dengxin Dai, and Luc Van Gool. Semantic foggy scene understanding with synthetic data. *International Journal of Computer Vision*, 126(9):973–992, 2018.
- [31] Rui Shao, Xiangyuan Lan, and Pong C Yuen. Feature constrained by pixel: Hierarchical adversarial deep domain adaptation. In *Proceedings of the 26th ACM international conference on Multimedia*, pages 220–228, 2018.
- [32] Zhiqiang Shen, Harsh Maheshwari, Weichen Yao, and Marios Savvides. Scl: Towards accurate domain adaptive object detection via gradient detach based stacked complementary losses. *arXiv preprint arXiv:1911.02559*, 2019.
- [33] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.
- [34] Peng Su, Kun Wang, Xingyu Zeng, Shixiang Tang, Dapeng Chen, Di Qiu, and Xiaogang Wang. Adapting object detectors with conditional domain normalization. *European Conference on Computer Vision*, 2020.
- [35] Jindong Wang, Wenjie Feng, Yiqiang Chen, Han Yu, Meiyu Huang, and Philip S Yu. Visual domain adaptation with manifold embedded distribution alignment. In *Proceedings of the 26th ACM international conference on Multimedia*, pages 402–410, 2018.
- [36] Tao Wang, Xiaopeng Zhang, Li Yuan, and Jiashi Feng. Few-shot adaptive faster r-cnn. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7173–7182, 2019.
- [37] Aming Wu, Yahong Han, Linchao Zhu, and Yi Yang. Instance-invariant domain adaptive object detection via progressive disentanglement. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021. doi: 10.1109/TPAMI.2021.3060446.
- [38] Rongchang Xie, Fei Yu, Jiaochao Wang, Yizhou Wang, and Li Zhang. Multi-level domain adaptive learning for cross-domain detection. *arXiv preprint arXiv:1907.11484*, 2019.
- [39] Chang-Dong Xu, Xing-Ran Zhao, Xin Jin, and Xiu-Shen Wei. Exploring categorical regularization for domain adaptive object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11724–11733, 2020.
- [40] Minghao Xu, Hang Wang, Bingbing Ni, Qi Tian, and Wen-jun Zhang. Cross-domain detection via graph-induced prototype alignment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12355–12364, 2020.
- [41] Fisher Yu, Wenqi Xian, Yingying Chen, Fangchen Liu, Mike Liao, Vashisht Madhavan, and Trevor Darrell. Bdd100k: A diverse driving video database with scalable annotation tooling. *arXiv preprint arXiv:1805.04687*, 2018.
- [42] Jiahui Yu, Yuning Jiang, Zhangyang Wang, Zhimin Cao, and Thomas Huang. Unitbox: An advanced object detection network. In *Proceedings of the 24th ACM international conference on Multimedia*, pages 516–520, 2016.
- [43] Ganlong Zhao, Guanbin Li, Ruijia Xu, and Liang Lin. Collaborative training between region proposal localization and classification for domain adaptive object detection. In *European Conference on Computer Vision*, pages 86–102. Springer, 2020.
- [44] Zhen Zhao, Yuhong Guo, Haifeng Shen, and Jieping Ye. Adaptive object detection with dual multi-label prediction. In *European Conference on Computer Vision*, pages 54–69. Springer, 2020.
- [45] Xinge Zhu, Jiangmiao Pang, Ceyuan Yang, Jianping Shi, and Dahua Lin. Adapting object detectors via selective cross-domain alignment. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 687–696, 2019.
- [46] Junbao Zhuo, Shuhui Wang, Weigang Zhang, and Qingming Huang. Deep unsupervised convolutional domain adaptation. In *Proceedings of the 25th ACM international conference on Multimedia*, pages 261–269, 2017.