remote sensing

MDPI

Article

Inter-Domain Invariant Cross-Domain Object Detection Using Style and Content Disentanglement for In-Vehicle Images

Zhipeng Jiang, Yongsheng Zhang, Ziquan Wang, Ying Yu *, Zhenchao Zhang, Mengwei Zhang, Lei Zhang and Binbin Cheng

School of Geospatial Information, PLA Strategic Support Force Information Engineering University, Zhengzhou 450001, China; jiangzp0803@163.com (Z.J.); yszhang2001@vip.163.com (Y.Z.); aresdrw@163.com (Z.W.); zhzhc_1@163.com (Z.Z); zhang295498@126.com (L.Z.); c18439912464@163.com (B.C.)
* Correspondence: yuying5559104@163.com; Tel.: +86-188-3715-3617

Abstract: The accurate detection of relevant vehicles, pedestrians, and other targets on the road plays a crucial role in ensuring the safety of autonomous driving. In recent years, object detectors based on Transformers or CNNs have achieved excellent performance in the fully supervised paradigm. However, when the trained model is directly applied to unfamiliar scenes where the training data and testing data have different distributions statistically, the model's performance may decrease dramatically. To address this issue, unsupervised domain adaptive object detection methods have been proposed. However, these methods often exhibit decreasing performance when the gap between the source and target domains increases. Previous works mainly focused on utilizing the style gap to reduce the domain gap while ignoring the content gap. To tackle this challenge, we introduce a novel method called IDI-SCD that effectively addresses both the style and content gaps simultaneously. Firstly, the domain gap is reduced by disentangling it into the style gap and content gap, generating corresponding intermediate domains in the meanwhile. Secondly, during training, we focus on one single domain gap at a time to achieve inter-domain invariance. That is, the content gap is tackled while maintaining the style gap, and vice versa. In addition, the style-invariant loss is used to narrow down the style gap, and the mean teacher self-training framework is used to narrow down the content gap. Finally, we introduce a multiscale fusion strategy to enhance the quality of pseudo-labels, which mainly focus on enhancing the detection performance for extreme-scale objects (very large or very small objects). We conduct extensive experiments on four mainstream datasets of in-vehicle images. The experimental results demonstrate the effectiveness of our method and its superiority over most of the existing methods.

Keywords: domain adaptation; object detection; domain gap disentanglement; inter-domain invariance

Citation: Jiang, Z.; Zhang, Y.; Wang, Z.; Yu, Y.; Zhang, Z.; Zhang, M.; Zhang, L.; Cheng, B. Inter-Domain Invariant Cross-Domain Object Detection Using Style and Content Disentanglement for In-Vehicle Images. *Remote Sens.* **2024**, *16*, 304. https://doi.org/10.3390/rs16020304

Academic Editors: Jong-Eun Ha, Hyoseok Hwang and Ronghui Zhan

Received: 9 November 2023
Revised: 29 December 2023
Accepted: 8 January 2024
Published: 11 January 2024

1. Introduction

Object detection (i.e., accurate classification and localization of objects of interest) based on optical images is a fundamental task [1] for remote sensing technology applied in autonomous vehicles. In recent years, due to the rapid development of deep learning, object detection methods has made huge breakthroughs. However, the performance mentioned above mainly obtained through training with labeled datasets under the fully supervised paradigm, but in real scenes, it is not feasible to always obtain labeled images. The variance of data will result in a significant decrease when these methods are applied to unfamiliar scenes. For autonomous driving, a remedy is to install many auxiliary sensors (such as LIDAR, etc.) to ensure the robust detection, but the cost is extremely high, far from reaching practicality.

Many studies [2–4] have shown that the domain shift between the training and the testing dataset primarily makes the performance degradation. To tackle this, many unsupervised domain adaptations (UDA) [5] have been used. In an UDA framework, the

labeled datasets often play as a "source" domain while the unlabeled one plays as the "target" domain; the method will try to transfer the knowledge from the source domain to the target domain. Thus, the UDA methods can maximize the utilization of both labeled/unlabeled data while mitigating the impact of domain discrepancy. Existing UDA methods can be divided into two categories: based on adversarial learning [4,6–15] and based on self-training [16–18]. Adversarial learning leverages GAN [19] or its variants to acquire domain-invariant features, enabling the transfer of detection knowledge, thereby enabling the model to perform well in the target domain. Self-training involves generating pseudo-labels directly using the model, then retraining the model to optimize the pseudo-labels, incrementally improving the model's adaptability. However, it is difficult for both of these methods to cross large domain gaps [20] due to the rough alignment strategy, for example, adversarial learning is unstable and easily obtains artifacts. Self-training methods have the same problem.

To tackle this problem and narrow the domain gap, some works [20,21] use intermediate domains to progressively implement unsupervised domain-adaptive object detection (UDA-OD) [22], but these methods mainly focus on how to reduce the style gap (e.g., the difference in color, texture, and brightness between the source and target domain) while ignoring the content gap (e.g., the variance in the traffic patterns and landscapes between the source and target domain). Inspired by previous research, we have discovered the ability to change image styles while preserving image content through stylization operations. Figure 1 illustartes our idea. CycleGAN [23], with its bi-directional nature, allows us to obtain intermediate domain images that possess distinct style and content gaps compared to the source domain images, simultaneously. Therefore, in this paper, we use CycleGAN to decouple the domain gap into the style and content gap, and generate a corresponding "synthetic domain" for subsequent model training. In order to make the model learn inter-domain-invariant features, we only deal with one gap at a time during the training process, adopting an alternating training strategy. For the style gap, we use the "style-invariant" loss to force the model to output the same prediction results under style changes. As for the content gap, we mainly adopt the mainstream mean teacher self-training [24] framework to enable the model to learn inter-domain content-invariant features. Finally, we transfer knowledge from the source domain to the target domain and make great improvement.

Our proposed method has the following four main contributions:

(1) Due to the negative relationship between the size of the domain gap and the performance of the detector, as well as the neglect of the content gap within the domain gap, we decouple the domain gap into the style and content gap to reduce the domain gap and generate the corresponding synthetic intermediate domain datasets.
(2) To ensure the effective learning in the inter-domain invariance, we employ alternating learning to separately handle the style and content gap. Additionally, we utilize style-invariant loss and the mean teacher self-training framework to address the style and content gap, respectively.
(3) We introduce a multiscale fusion strategy to enhance the detection performance of extreme-scale (very large or very small) objects, thereby improving the quality of pseudo-labels.
(4) Through comprehensive experiments conducted on various adaptation benchmarks in the context of autonomous driving scenarios, we have demonstrated that our proposed method outperforms the majority of existing methods.

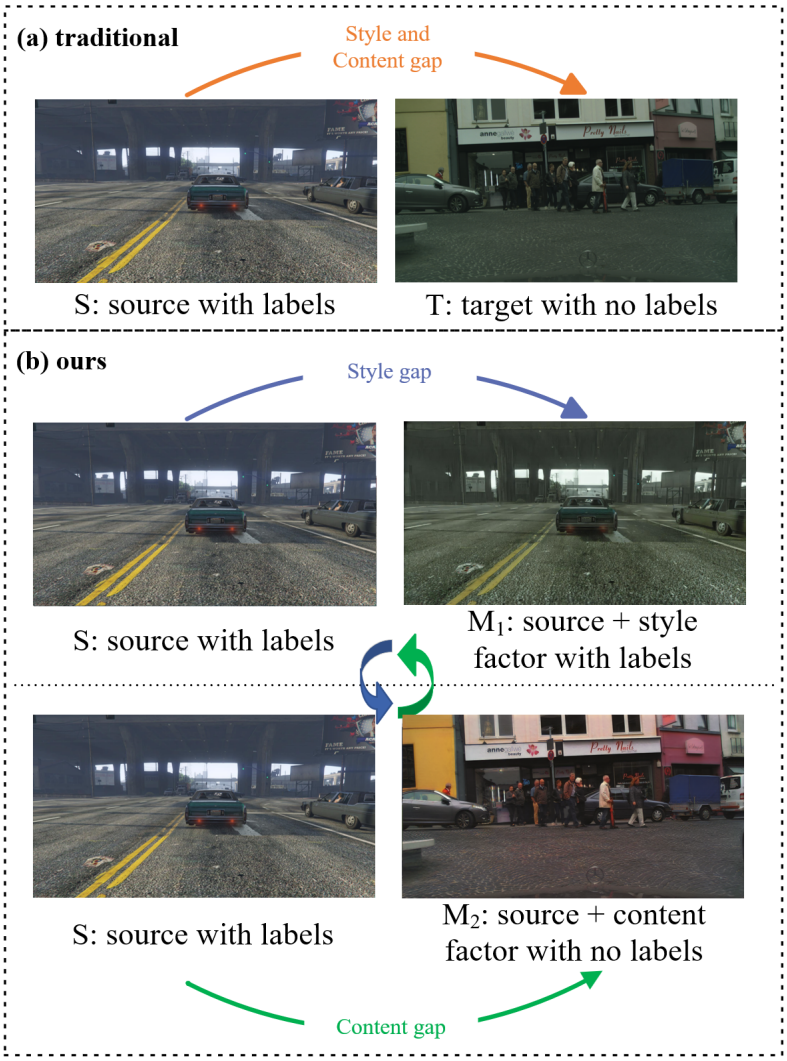

Figure 1. Main idea of traditional method and our method. (**a**) In most traditional methods, both advertised training and self-training directly handle the domain gap to achieve cross domain object detection. (**b**) Due to the adverse effect of domain gap on the performance of cross-domain object detection, we employ CycleGAN [23] to decouple the domain gap into the style and content gap, thereby reducing the domain gap. At the same time, in order to ensure inter-domain invariance, we adopt alternating training to separately handle the style and content gap to avoid interference between the two information about the style and content factor.

2. Related Work

2.1. Object Detection

The main task of object detection is to locate the bounding boxes of objects and determine their categories. Relevant research can be divided into traditional methods [25–28] and deep learning-based methods [29–43]. Traditional object detection algorithms are essentially sliding window + traditional machine learning classifiers, which mainly rely on handcrafted algorithms and the corresponding generated features. However, with the advent of big data and the era of deep learning in recent years, significant breakthroughs have been made in object detection. Currently, object detection models are mainly divided into two categories: two-stage and one-stage. In two-stage object detection algorithms, the primary ones are the RCNN series algorithms [29–32], which consist of two steps: generating candidate regions, extracting features from these regions and performing object classification and bounding box regression. Due to the limitations of the two-stage approach, this method inevitably leads to lower detection efficiency. On the other hand, one-stage object

detection techniques effectively address this issue. This category of algorithms includes the YOLO series [33–36] and SSD [37,38], among others.

In recent years, after the application of ViT [39] in object detection, transformer-based detection algorithms have emerged. These algorithms can mainly be divided into two series: Transformer Neck and Transformer Backbone, which focus on replacing CNN-based Neck (Backbone) with Transformer-based Neck (Backbone). For Transformer Neck, the main works are DETR [40] and its variant Deformable DETR [41]. For Transformer Backbone, the most significant works are Swin Transformer [42] and PVT (Pyramid Vision Transformer) [43]. These works mainly focus on how to better extract image feature using Transformer. Due to the state-of-the-art performance of the Swin Transformer algorithm and its variants on the COCO [44] dataset, we choose Swin Transformer [42] as our baseline detector to improve its cross-domain detection performance in autonomous driving scenarios.

2.2. Unsupervised Domain Adaptive Object Detection (UDA-OD)

UDA-OD is an effective method that uses the UDA approach to solve cross-domain object detection from the labeled source domain to the target domain. Among these approaches, they can be mainly divided into two categories: adversarial learning and self-training. Adversarial learning approaches primarily utilize GAN [19] to align the features between the source and target domains, aiming to achieve cross domain object detection. Domain-adaptive faster RCNN (DAF) [4] is the first method that utilizes adversarial training to address unsupervised domain adaptive object detection by aligning features. Subsequent works based on adversarial learning differ mainly in the location of feature alignment: image-instance feature alignment [6–8], global-local feature alignment [9–12], and object region feature alignment [13–15].

In recent years, studies have shown that self-training is more stable and effective compared to adversarial training. One of the most prominent approaches is utilizing the mean teacher [24] self-training framework, which extends the idea from semi-supervised object detection to UDA-OD. MTOR [16] utilizes the mean teacher framework to consider consistency at both the region level and graph structure level, enabling cross-domain object detection. Unbiased mean teacher (UMT) [17] reduces the student and teacher model bias by combining CycleGAN with mean teacher [24]. Despite the performance improvement achieved by MTOR and UMT, during the training process, the pseudo-labels generated by the teacher model still have a certain error rate due to domain shift. Ref. [20] proposed utilizing CycleGAN to stylize the source domain and generate intermediate domain with the style of the target domain. By leveraging the intermediate domain and the target domain, the aim is to effectively reduce the domain gap and subsequently perform cross-domain object detection. Adaptive teacher (AT) [18] combines adversarial learning with the mean teacher framework to narrow the domain gap through adversarial learning in the student model, thereby enhancing the quality of pseudo-labeling and further improving model performance. Contrastive mean teacher (CMT) [45] delves into the intrinsic connection between mean teacher self-training and contrastive learning. Based on this, it naturally integrates the two paradigms of contrastive learning and mean teacher, aiming to maximize beneficial learning signals. This method uses pseudo-labels to extract object-level features and optimizes them through contrastive learning without requiring labeling in the target domain. Owing to the popularity of DETR-style detectors, sequence feature alignment (SFA) [14] has introduced a cross-domain detector based on deformable DETR. O^2net [46] also proposed an end-to-end detector based on multilevel feature alignment and a mean teacher framework.

While the aforementioned generation of intermediate domain images effectively reduces the domain gap, they only utilize labeled source domain images with the style of the target domain, overlooking the unlabeled source domain images with the style of the target domain. In this paper, we consider two intermediate domains that have both style and content gaps from the source domain. We employ alternating training to address these two differences separately, thereby preserving inter-domain invariance.

2.3. Pseudo-Label Optimization

In the mean teacher self-training framework, one crucial aspect for cross-domain performance of models is the quality of pseudo-labels. Therefore, improving the quality of pseudo-labels has become a research focus for many scholars. PCdKD [47] proposes a two-stage strategy that combines domain adaptation and knowledge distillation, gradually achieving feature-level adaptation to obtain reliable pseudo-labels. SC-UDA [48] utilizes uncertainty-based detection pseudo-labeling to obtain better pseudo-labels for training. Mixteacher [49] generates high-quality pseudo-labels using a mixed-scale feature pyramid. Inspired by Mixteacher [49], we are giving more attention to extreme scale targets, i.e., very large or very small size targets. Therefore, in this paper, we apply a multiscale fusion strategy to optimize the pseudo-label generation process in the teacher model.

3. Method

3.1. Overview

In the context of UDA-OD problem, we require a source domain $\mathcal{S}$ and a target domain $\mathcal{T}$, where $\mathcal{S}$ contains N_s labeled images $D_s = \{(X_s, Y_s)\}$, and $\mathcal{T}$ contains T_s unlabeled images $D_t = \{X_t\}$. X_s and Y_s represent the image and its corresponding label in the source domain, while X_t represents the image in the target domain. Due to the negative correlation between the size of the domain gap between the source and target domains and the performance of cross-domain detector mentioned above, we utilize CycleGAN [23] to decouple the domain gap into the style and content gap, and simultaneously generate two corresponding synthetic intermediate domains ($\mathcal{M}_1$ and $\mathcal{M}_2$).

The main structure of the IDI-SCD proposed in this paper is shown in Figure 2. Regarding the models used during training, we primarily establish two models: the student model F_s and the teacher model T_t. The student model serves as the main model and is utilized for supervised training in the source domain and style-invariant training to adress the style gap. Additionally, it forms the Mean Teacher self-training framework along with the teacher model to handle the content gap. During the model training, it is essential to ensure inter-domain invariance, meaning that the content varies while the style remains the same, and vice versa. To achieve this, we employ alternating training, ensuring that each iteration during training only one gap is addressed, thus avoiding mutual interference during knowledge transfer.

3.2. Supervised Training in the Source Domain

In the mean teacher self-training framework, it is of utmost importance to ensure that the teacher model obtains reliable pseudo-labels at the beginning of training. To achieve this goal, we employ supervised training and initialization of the student model using labeled images $D_s = \{(X_s, Y_s)\}$ from the source domain. The supervised loss can be defined as follows:

$$\mathcal{L}_{sup}(X_s, Y_s) = \mathcal{L}_{cls}(X_s, Y_s) + \mathcal{L}_{reg}(X_s, Y_s) \tag{1}$$

where $\mathcal{L}_{cls}(X_s, Y_s)$ and $\mathcal{L}_{reg}(X_s, Y_s)$ represent the classification loss and bounding box regression loss, respectively.

3.3. Style Invariant Loss

Through the stylization operation of CycleGAN, we have discovered that each image inputted into the student model possesses two styles: the source domain style and the target domain style. Most methods [20,21] only utilize the source domain images with the target domain style while disregarding the the images in the source domain. However, when it comes to cross-domain object detection, it is crucial to achieve output invariance for images with different styles but identical content. In this paper, We ensure output invariance by maintaining the invariance of features generated by Feature Encoder (Backbone) for both $\mathcal{S}$ and $\mathcal{M}_1$, thereby enabling the model to maintain invariance for images with the same content as the input.

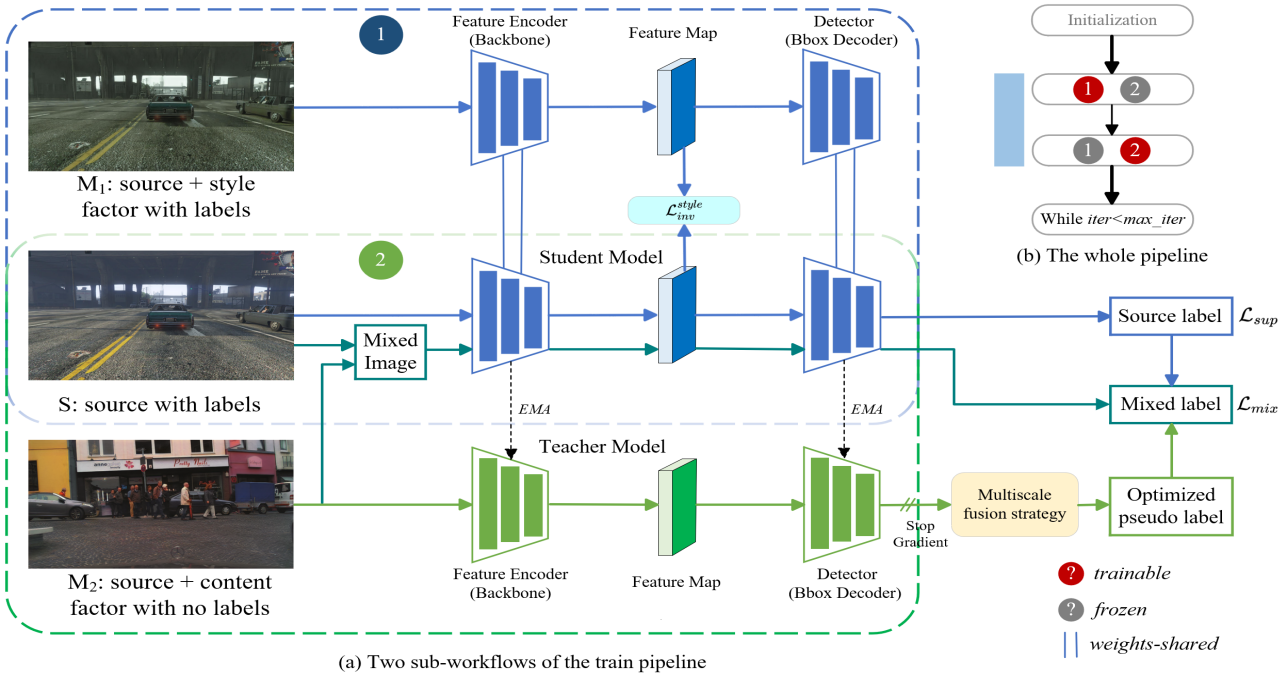

Figure 2. The proposed method. (**a**) By decoupling the inter-domain gap into the style and content gap, we simultaneously obtain two datasets that have distinct style and content gap from the source domain. To address the style and content gap, we have designed two separate sub-workflows specifically tailored to handle each of them. For style gap, we utilize the style-invariant loss to ensure that the student model maintains feature invariance when there is only a change in style in the input image, thereby ensuring consistency in the output results. On the contrary, we employ the mean teacher self-training framework to address content gap, while utilizing a multiscale fusion strategy to obtain higher-quality pseudo-labels. (**b**) To maintain inter-domain invariance and prevent information interference during the handling of different domain gaps, we train the two sub-workflows alternately. Additionally, to ensure effective gradient backpropagation, we only perform gradient backpropagation on the student model, while updating the parameters of the teacher model using exponential moving average (EMA) based on the student model's parameters.

More specifically, we divide the model into two parts: Feature Encoder (Backbone) g and Detector (Bbox Decoder) h. For two input images X_s and X_{m_1} with the same size and different style, the output feature representations of the model are denoted as $g(X_s)$ and $g(X_{m_1})$. We define the feature invariance loss, also known as style-invariant loss, as follows:

$$\mathcal{L}_{inv}^{style}(g(X_s), g(X_{m_1})) = \frac{1}{K}\sum_{i=1}^{K}\frac{1}{C_i H_i W_i}||g_i(X_s) - g_i(X_{m_1})||_2^2 \quad (2)$$

where g_i is output of the ith layer of the Feature Encoder (Backbone) $g(x)$ when processing the image x, which is a feature map of shape $C_i \times H_i \times W_i$. $||*||_2^2$ represents the Euclidean distance.

3.4. Self Training in the Target Domain

As obtaining real labels for supervised training in the target domain is not feasible, we utilize the pseudo-labels generated in the target domain to train the model. However, relying solely on pseudo-labels for training may lead to deviation in the model's prediction results due to their quality. Therefore, in this paper, we adopt the ConfMix [50] to mix the source domain and the target domain, which imposes certain constraints on the model using the real labels of the source domain, achieving a similar "semi-supervised" purpose.

To be specific, first, the image X_t from the target domain is input into the teacher model to obtain a series of classification scores for candidate foreground objects, as well as their corresponding bounding box coordinates. Next, the scores that are greater than a threshold

τ are retained to obtain the final pseudo-labels $\widetilde{Y}_t$. In this step, we utilize a multiscale fusion strategy to improve the quality of the pseudo-labels. Finally, based on the scores of the pseudo-labels, the average score of each region (here, each image is divided into four regions) is calculated to determine which area will be blended with the source domain image. As a result, we can obtain the mixed image X_{mix} along with its corresponding labels Y_{mix}. With this setup, we train the model with X_{mix} and Y_{mix}; the loss $\mathcal{L}_{mix}$ can be defined accordingly:

$$\mathcal{L}_{mix}(X_{mix}, Y_{mix}) = \mathcal{L}_{cls}(X_{mix}, Y_{mix}) + \mathcal{L}_{reg}(X_{mix}, Y_{mix}) \tag{3}$$

As illustrated in Figure 2 ②, the gradient backpropagation of the teacher model is frozen. However, as the training progresses, we need to update the teacher model parameters to obtain better pseudo-labels, so we employ exponential moving average (EMA) to update the parameters of the teacher model using a subsequent formula. Specifically, during training, the parameters of current teacher model is the linear sum of the parameters of previous student and teacher model from the previous training iteration.

$$\phi_{t+1} \leftarrow \alpha\phi_t + (1-\alpha)\theta_t \tag{4}$$

where θ_t and ϕ_t represent the parameters of the student and teacher models at time t, respectively, and ϕ_{t+1} represents the parameters of the teacher model at time $t+1$. In the context of t and $t+1$, they denote the previous iteration and the current iteration during model training. The smoothing coefficient hyperparameter is denoted as α.

3.5. Pseudo-Label Optimization Based on Multi-Resolution

In previous methods [17,18], resized images were input into the teacher model to obtain pseudo-labels. However, images at that size are not able to effectively detect extremely large or small objects. Therefore, in this paper, we perform upsampling and downsampling on the images separately, making them suitable for detecting extremely small or large targets and propose the multiscale fusion strategy. The detail is illustrated in Figure 3.

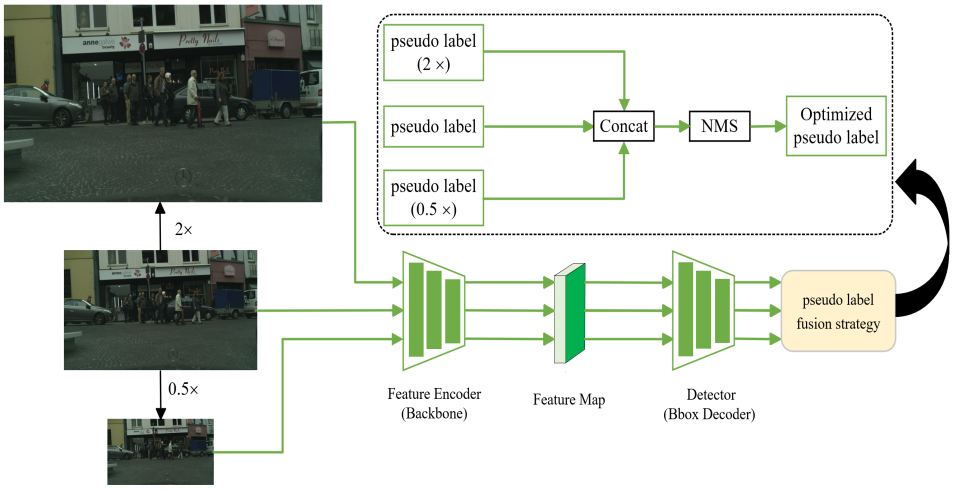

Figure 3. Illustration of the multiscale fusion strategy. When generating pseudo-labels with the teacher model, we simultaneously input three images of different resolutions and generate three sets of pseudo-labels. Next, we will perform the inverse operation on the coordinates of the pseudo-labels generated from the upsampled and downsampled images, in order to map the coordinates back to the original image's coordinate space. Finally, we merge the three sets of processed pseudo-labels and perform non-maximum suppression (NMS) to obtain the optimized pseudo-labels.

Specifically, firstly, we upsample and downsample the images and input them together with the original image into the teacher model, obtaining three sets of pseudo-labels. Next, we will map all the coordinates from three sets of pseudo-labels back to the original image's coordinate space. Finally, we merge the three sets of pseudo-labels and filter the merged pseudo-labels using NMS to obtain the final pseudo-labels.

3.6. Total Loss

As illustrated in Figure 2, We utilize the alternation of ① and ② for training to separately address the style and content gap. For the alternating training, the models are trained with ① during odd iterations ($l = 2n + 1$) and ② during even iterations ($l = 2n$). Meanwhile, we utilize EMA to update the parameters of the teacher model in each iteration.

As a result, the total loss can be calculated as follows:

$$\mathcal{L}_{total} = \begin{cases} \mathcal{L}_{sup}(X_s, Y_s) + \lambda_1 \mathcal{L}_{mix}(X_{mix}, Y_{mix}) & l = 2n \\ \mathcal{L}_{sup}(X_s, Y_s) + \lambda_2 \mathcal{L}_{inv}^{style}(g(X_s), g(X_{m_1})) & l = 2n + 1 \end{cases} \tag{5}$$

where λ_1 is a dynamically changing hyperparameter based on the quality of the pseudo-labels and λ_2 is a tunable hyperparameter.

4. Experimental Result

4.1. Datasets

In our experiments, we use four public datasets: Sim10k [51], BDD100K [52], Cityscapes [53] and Kitti [54] to evaluate our proposed method.

Sim10k [51] is an synthetic dataset rendered directly by the GTA5 game engine, consisting of 10,000 images and 58,701 corresponding car annotations.

Kitti [54] is a real-world dataset collected in various road scenes, which comprises 7481 training images. The annotations of this dataset mainly consist of eight classes related to road targets. In this paper, we only use the "car" category to evaluate the performance of the model in cross-camera adaptation.

BDD100K [52], a large-scale autonomous driving dataset with various time periods, diverse weather conditions (including sunny, cloudy, and rainy, as well as different times of day such as daytime and evening), and driving scenes. In our experiment, we select and obtain images captured under sunny weather conditions to validate the adaptability of our model to scene variations.

Cityscapes [53] is a dataset captured in urban autonomous driving scenarios, including 2975 images for training and 500 images for validation.

Based on the aforementioned four datasets, we can obtain two cross-domain scenarios under autonomous driving conditions: (1) synthetic to real adaptation (Sim10k → Cityscapes), where the images in the source domain and target domain are captured from the synthetic and real-world scenarios; (2) across-cameras adaptation (Cityscapes → BDD100k and KITTI → Cityscapes), where all the three datasets are taken from various cities using different cameras. The combination of all datasets and cross-domain scenarios is presented in Table 1.

Table 1. The combination of different cross domain scenarios. *S*, *C*, *K* and *B* represent Sim10k, Cityscapes, KITTI, and BDD100k, respectively.

Cross Domain Scenarios	Trainig Set		Validation Set
	Source Domain	Target Domain	Target Domain
$S \rightarrow C$	KITTI	Cityscapes	Cityscapes
	10,000	2975	500
$K \rightarrow C$	Sim10k	Cityscapes	Cityscapes
	7481	7481	500
$C \rightarrow B$	Cityscapes	BDD100k	BDD100k
	2975	36728	5258

4.2. Implementation Details

In this paper, RetinaNet [38] serves as the benchmark detector, and the original ResNet is replaced with Swin-T as the new backbone. During training, the student model is trained by the SGD [55] with a learning rate of 0.00125 and a batch size of 2. Moreover, in order to make model training more stable, we implement a linear learning rate warmup in the first 0.5k iteration. Regarding the image size, We followed [18] to adjust the shorter side of all images to 600 and keep the image aspect ratio constant. For the hyperparameters in the mean teacher self-training framework, we set the score threshold τ for pseudo-label filtering to 0.7, and the smoothing coefficient α for EMA to 0.999. The total number of iterations is set to 80k. The default value λ_2 of the weights in Equation (5) is set to 10 for synthetic-to-real adaptation, and 5 for across-cameras adaptation. All experiments were conducted using the mmdetection framework [56] and PyTorch [57].

4.3. Performance Comparison

Synthetic-to-Real Adaptation. In previous research, we have typically collected data manually and annotated it artificially. However, this approach is time-consuming, labor-intensive, and costly. With the advancement of data synthesis engines, it is now possible to automatically generate synthetic datasets and corresponding labels based on actual needs. Therefore, it is highly worthwhile to investigate how models trained on synthetic datasets can be adapted to real-world scenarios. In this setting, we utilize the Sim10k dataset as the source domain and Cityscapes as the target domain to evaluate the cross-domain performance of our proposed method in the synthetic-to-real adaptation. We compare several state-of-the-art methods, and the experimental results are shown in Table 2. From the experimental results, we can observe that our proposed method achieve an improvement of +1.1 compared to the state-of-the-art method OADA [58] in the car category.

Table 2. Performance Comparison I. Comparative results of different unsupervised domain adaptation methods on Sim10k → Cityscapes benchmark (car AP).

Method	Detector	Backbone	Car AP
Source	RetinaNet	Swin-T	41.5
DAF [4]	Faster R-CNN	VGG-16	41.9
SWDA [9]	Faster R-CNN	ResNet-101	44.6
SCDA [13]	Faster R-CNN	VGG-16	45.1
MTOR [16]	Faster R-CNN	ResNet-50	46.6
CR-DA [59]	Faster R-CNN	VGG-16	43.1
CR-SW [59]	Faster R-CNN	VGG-16	46.2
SAD [60]	Faster R-CNN	ResNet-50	49.2
ViSGA [8]	Faster R-CNN	ResNet-50	49.3
D-adapt [61]	Faster R-CNN	ResNet-101	53.2
EPM [62]	FCOS	ResNet-101	47.3
SIGMA [63]	FCOS	ResNet-50	53.7
MGA [64]	FCOS	ResNet-101	55.4
OADA [58]	FCOS	VGG-16	56.6
SFA [14]	D-DETR	ResNet-50	52.6
O^2net [46]	D-DETR	ResNet-50	54.1
Our (IDI-SCD)	RetinaNet	Swin-T	**57.7**

Note that, the best car AP is highlighted in bold format.

Across Cameras Adaptation. On the other hand, due to the thriving development of autonomous driving, numerous scholars have proposed different road object detection datasets. These datasets are obtained from various cities using different cameras, resulting in significant variations in terms of style, resolution, and other aspects of the captured images. Such variations can negatively affect the performance of trained detectors in practical

applications, so we need to make sure that our proposed approach is able to mitigate the domain shift in different datasets so as to ensure robust performance and generalization. And in order to compare it with existing state-of-the-art methods, we have chosen three public datasets: Cityscapes, KITTI, and BDD100k. We conduct two experiments in across-cameras adaptation: KITTI → Cityscapes and Cityscapes → BDD100k. The results of these two experiments are shown in Tables 3 and 4, respectively. For KITTI → Cityscapes, our model receives +0.6 improvement in the "car" category. Moreover, a performance improvement of +1.7 is achieved in Cityscapes → BDD100k.

Table 3. Performance Comparison II. Comparative results of different unsupervised domain adaptation methods on KITTI → Cityscapes benchmark (car AP).

Method	Detector	Backbone	Car AP
Source	RetinaNet	Swin-T	42.3
DAF [4]	Faster R-CNN	VGG-16	41.8
SWDA [9]	Faster R-CNN	ResNet-101	43.2
RKTMG [15]	Faster R-CNN	ResNet-50	43.5
SCDA [13]	Faster R-CNN	VGG-16	43.6
ViSGA [8]	Faster R-CNN	ResNet-50	47.6
EPM [62]	FCOS	ResNet-101	45.0
SIGMA [63]	FCOS	ResNet-50	45.8
OADA [58]	FCOS	VGG-16	46.3
MGA [64]	FCOS	ResNet-101	47.6
MS-DAYOLO [65]	YOLOv5	CSP-Darknet53	47.6
DAYOLO [66]	YOLOv5	CSP-Darknet53	48.7
S-DAYOLO [66]	YOLOv5	CSP-Darknet53	49.3
Our (IDI-SCD)	RetinaNet	Swin-T	**49.9**

Note that, the best car AP is highlighted in bold format.

Table 4. Performance Comparison III. Comparative results of different unsupervised domain adaptation methods on Cityscapes → BDD100k benchmark (mAP).

Method	Detector	Backbone	Person	Rider	Car	Truck	Bus	Mcycle	Bicycle	mAP
Source	RetinaNet	Swin-T	35.0	27.4	54.7	17.3	11.2	15.1	22.8	26.2
DAF [4]	Faster R-CNN	VGG-16	28.9	27.4	44.2	19.1	18.0	14.2	22.4	24.9
SWDA [9]	Faster R-CNN	ResNet-101	29.5	29.9	44.8	20.2	20.7	15.2	23.1	26.2
SCDA [13]	Faster R-CNN	VGG-16	29.3	29.2	44.4	20.3	19.6	14.8	23.2	25.8
CR-DA [59]	Faster R-CNN	VGG-16	30.8	29.0	44.8	20.5	19.8	14.1	22.8	26.0
CR-SW [59]	Faster R-CNN	VGG-16	32.8	29.3	45.8	**22.7**	20.6	14.9	25.5	27.4
EPM [62]	FCOS	ResNet-101	39.6	26.8	55.8	18.8	19.1	14.5	20.1	27.8
SFA [14]	D-DETR	ResNet-50	40.2	27.6	57.8	19.1	23.4	15.4	19.2	28.9
O^2net [46]	D-DETR	ResNet-50	40.4	31.2	58.6	20.4	**25.0**	14.9	22.7	30.5
Our (IDI-SCD)	RetinaNet	Swin-T	**43.5**	**34.0**	**61.1**	18.4	15.6	**21.1**	**32.3**	**32.2**

Note that, the best AP in each category and mAP are highlighted in bold format.

4.4. Discussion

The optimal value of the hyperparameter λ_1 and λ_2. First, we discuss the optimal value of λ_1. During the training process, there are differences in the quality of pseudo-labels obtained from different images. Based on this, we apply a dynamically changing hyperparameter λ_1 to $\mathcal{L}_{mix}$, which is proportional to the quality of pseudo-labels. To validate the effectiveness of λ_1, we conducted comparative experiments by setting λ_1 to different constant values of 0.2, 0.4, 0.6, 0.8, and 1. The experimental results are shown in Table 5. We found that utilizing λ_1 achieves the best detection performance, with a performance improvement of 0.7 compared to when the weight is set to a constant value

0.6. This well demonstrates that dynamically adjusting the proportion of $\mathcal{L}_{mix}$ based on the quality of pseudo-labels can allow the model to achieve the optimal performance. Now, we compare the the default value 10 of λ_2 with five other values. As depicted in Table 6, when λ_2 = 10, the model has the best performance. However, we also observed that when λ_2 changes, the car AP fluctuates within a small range, indicating a weak correlation between the optimal performance that the model can achieve and λ_2. Based on the experimental process, it is highly likely due to the sufficiently large total number of iterations, which allows the model to fully learn relevant knowledge about the style factor, resulting in minimal fluctuations in model performance.

Table 5. Ablation study on hyperparameter λ_1. All the experiments are conducted on the Cityscapes-value dataset. We compare the λ_1 with five other constant weights.

Weight	**0.2**	**0.4**	**0.6**	**0.8**	**1.0**	$\boldsymbol{\lambda_1}$
car AP	52.1	56.3	57.0	56.7	51.2	**57.7**

Note that, the best car AP is highlighted in bold format.

Table 6. Ablation study on hyperparameter λ_2. All the experiments are conducted on the Cityscapes-value dataset. We compare the default value 10 with five other values.

$\boldsymbol{\lambda_2}$	**1**	**2**	**5**	**10**	**15**	**20**
car AP	57.3	57.6	57.0	**57.7**	57.5	57.4

Note that, the best car AP is highlighted in bold format.

Effectiveness of the multiscale fusion strategy. As mentioned above, in the mean teacher self-training framework, improving the quality of pseudo-labels is the most effective method to enhance cross-domain knowledge transfer. Typically, we only obtain corresponding pseudo-labels using single-resolution images. Therefore, we propose a multiscale fusion strategy. To validate the effectiveness of our proposed module, we conducted a series of ablation experiments on the Cityscapes dataset. The experimental results are shown in Table 7. With the addition of low-resolution images, there is a significant improvement of +1.5 in car AP for large objects compared to the absence of such images. Moreover, by solely incorporating high-resolution images, there is a notable enhancement of +2.2 in car AP for small objects. These findings further reinforce the positive impact of integrating high-resolution or low-resolution images on performance enhancement while ensuring that the detection performance of objects of other sizes is either maintained or even slightly improved. Finally, by simultaneously combining pseudo-labels obtained from high-resolution and low-resolution images, we achieve a "1 + 1 > 2" effect, whereby the detection performance of objects of different sizes has been further improved compared to solely adding a single resolution image. This validates the effectiveness of the multiscale fusion strategy in enhancing the detection performance of large and small objects.

Table 7. Analysis of multiscale fusion strategy. We conduct different combinations of the fusion strategies on Cityscapes-value datasets. 1.0×, 0.5× and 2× indicate the image with original size, downsampled to 0.5× and upsampled to 2×, respectively.

Different Combinations of Image Inputs			**Car AP**	**Car $\mathbf{AP_s}$**	**Car $\mathbf{AP_m}$**	**Car $\mathbf{AP_l}$**
1.0×	**0.5×**	**2×**				
✓			55.2	24.2	66.2	86.0
✓	✓		55.8	25.5	66.0	87.5
✓		✓	56.3	26.4	66.9	87.3
✓	✓	✓	**57.7**	28.0	68.8	88.5

Note that, the best car AP is highlighted in bold format.

Comparison of methods for dealing with the domain gap. Table 8 illustrates the results of different methods for dealing with the domain gap. Due to our decoupling of the domain gap into the style gap and content gap, we conducted a series of ablation experiments to validate the superiority of our proposed method. Currently, we primarily encounter the style gap, content gap, and the genuine domain gap that arises from the combination of both in the source and target domains. Therefore, we first conducted experiments on the style gap, content gap, and domain gap between $\mathcal{S}$ and $\mathcal{M}_1$, $\mathcal{S}$ and $\mathcal{M}_2$, $\mathcal{S}$ and $\mathcal{T}$. Based on the experimental results shown in Table 2, we observed that among the three methods, $\mathcal{S} \rightarrow \mathcal{M}_2$ achieved the best performance, followed by $\mathcal{S} \rightarrow \mathcal{T}$. This can be attributed to two main factors: (1) Compared to $\mathcal{S} \rightarrow \mathcal{T}$, $\mathcal{S} \rightarrow \mathcal{M}_2$ reduces the domain gap, resulting in improved quality of the pseudo-label obtained by the teacher model in the mean teacher self-training framework, thereby enhancing cross-domain detection performance. (2) Compared to $\mathcal{S} \rightarrow \mathcal{T}$, although the style gap handled between $\mathcal{S}$ and $\mathcal{M}_1$ have significantly reduced compared to the original gap between $\mathcal{S}$ and $\mathcal{T}$ in actual domains, we also discovered that this will result in quite small gap between $\mathcal{S}$ and $\mathcal{M}_1$. At this moment, when we utilize style-invariant loss for style knowledge learning, even if we can fully learn the relevant knowledge, the cross-domain performance achieved cannot reach the level of directly using $\mathcal{S} \rightarrow \mathcal{T}$. In summary, we find that both the style gap and the content gap play a significant role in improving the cross-domain performance of the model. From the last two rows of Table 8, we find that dealing with both of the two gaps can effectively promote the performance improvement of cross-domain detectors. In particular, the use of alternate learning to deal with the style and content gap separately can achieve the optimal performance. This is because when dealing with two domain gaps at the same time, the relevant knowledge transferred can interfere with each other. Therefore, in this paper, we use the alternating training strategy to deal with style and content gaps, respectively. The specific details of the performance enhancement achievable through this training strategy will be further elucidated in the subsequent section, where visualized results will provide a more comprehensive explanation.

Qualitative results. Based on the various quantitative analyses provided above, in order to provide a more intuitive visualization and facilitate discussion, we visualized and compared the model's predicted results. The intuitive results are shown in Figures 4 and 5.

Table 8. Combination of two sub-workflows in train pipeline. All the experiments are conducted on Cityscapes-value dataset. And the multiscale fusion strategy has been used.

Combination of Two Sub-Workflows in Train Pipeline		w/AT [1]	mAP
$\mathcal{S} \rightarrow \mathcal{M}_1$			48.3
$\mathcal{S} \rightarrow \mathcal{M}_2$			53.9
$\mathcal{S} \rightarrow \mathcal{T}$			52.7
$\mathcal{S} \rightarrow \mathcal{M}_1$	$\mathcal{S} \rightarrow \mathcal{M}_2$		55.2
		✓	**57.7**

[1] AT means alternate training. The best car AP is highlighted in bold format.

As shown in Figure 4, we compared the results of four model training approaches: source, + $\mathcal{S} \rightarrow \mathcal{M}_1$, + $\mathcal{S} \rightarrow \mathcal{M}_2$ and our proposed IDI-SCD, against the ground truth labels. We observe that without cross-domain model training, there are significant instances of duplicate detections, missed detections, and false detections in the predicted results. However, with the inclusion of either $\mathcal{S} \rightarrow \mathcal{M}_1$ or $\mathcal{S} \rightarrow \mathcal{M}_2$ (e.g., dealing with the style gap or the content gap). there was a substantial reduction in duplicate detections and false detections within the model. Furthermore, by combining both approaches and effectively addressing the both the style and content gap, there is further improvement in detection performance for overlapping or small objects. For more detailed information about these results, please refer to Figure 5.

Figure 4. Qualitative performance of ablation study in our method. These experiments are conducted on the Cityscapes-value dataset. It demonstrates the improvement in cross-domain detection performance achieved by addressing the style or the content gap in our proposed methods, leading to the optimal performance when these two components are integrated.

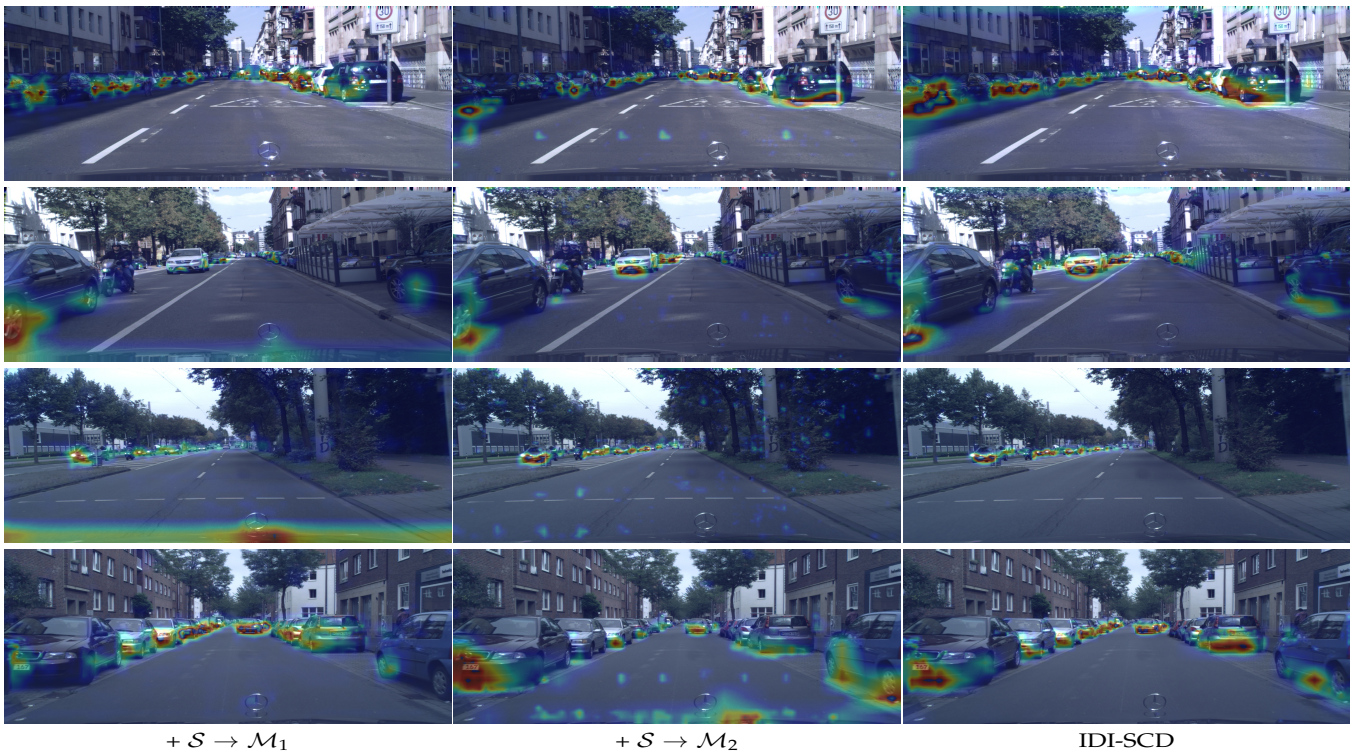

Figure 5. Heatmap visualization results of ablation study in our method. These experiments are conducted on the Cityscapes-value dataset. This provides another representation of the results in Figure 4, showcasing additional details.

5. Conclusions

In this paper, our aim is to improve the performance of the cross-domain domain object detector in autonomous driving across domain scenarios. Instead of focusing on using adversarial learning to align features or improve the quality of pseudo-labels in self-training frameworks, we primarily focus on how to build a simple, efficient, and plug-and-play plugin. As a result, We first discuss the relationship between the size of the domain gap and the performance of the detector and find that they are negatively correlated. Based on the aforementioned finding, we need the intermediate domain to narrow the domain gap. To achieve this, we decouple the domain gap into the style and content gap using CycleGAN and generate the corresponding synthetic intermediate domain datasets for model training. In the meantime, the effectiveness in narrowing the domain gap is validated through a series of experiment. Due to the reduction in domain discrepancy, we decompose the original cross-domain knowledge transfer into two steps, focusing on either style or content differences during each training iteration. Based on this, we propose a novel unsupervised domain adaptive object detection method named IDI-SCD in autonomous driving cross domain scenarios. In addition, we propose the multiscale fusion strategy to enhance the detection performance of the model for extreme scale objects, specifically those with either very large or very small sizes.

Finally, our proposed method has been tested on two cross-domain scenarios (three domain adaptation benchmarks), which demonstrates that our method achieves certain performance improvements compared to current mainstream methods. Specifically, our proposed method achieves an improvement of +1.1 and +0.6 compared to the state-of-the-art method MGA in the car category on Sim10k $\rightarrow$ Cityscapes and KITTI $\rightarrow$ Cityscapes, respectively. Moreover, on Cityscapes $\rightarrow$ BDD100k, it also increases the performance by O^2net by 1.7 in terms of mAP.

Due to the certain limitations in decoupling the domain gap with CycleGAN, we will investigate how to better decouple the domain gap and achieve better transfer of cross-domain knowledge in the future. Additionally, we will conduct more in-depth research on methods to improve the quality of pseudo-labels. We will also study universal object detectors applicable to various scenarios.

Author Contributions: Conceptualization, Z.J. and Z.W.; methodology, Z.J., Y.Y. and Z.W.; software, Z.J. and Z.W.; validation, Z.J., Y.Z. and Y.Y.; formal analysis, Y.Z., Y.Y. and B.C.; investigation, M.Z., Z.Z., Y.Y., B.C. and L.Z.; data curation, Z.J., Z.W., Y.Y., M.Z. and L.Z.; writing—original draft preparation, Z.J., Z.W. and L.Z; writing—review and editing, Z.J., Z.W., Z.Z., M.Z. and L.Z.; visualization, Y.Z. and Z.Z; supervision, Y.Z. and Y.Y.; project administration, Y.Z.; funding acquisition, Y.Z. and Y.Y. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the National Natural Science Foundation of China under Grant 42071340 and Program of Song Shan Laboratory (included in the management of Major Science and Technology of Henan Province) under Grant 2211000211000-01.

Data Availability Statement: The data presented in this study are publicly available on the corresponding official website.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Zhu, H.; Yuen, K.V.; Mihaylova, L.; Leung, H. Overview of environment perception for intelligent vehicles. *IEEE Trans. Intell. Transp. Syst.* **2017**, *18*, 2584–2601. [CrossRef]
2. Gopalan, R.; Li, R.; Chellappa, R. Domain adaptation for object recognition: An unsupervised approach. In Proceedings of the 2011 International Conference on Computer Vision, Barcelona, Spain, 6–13 November 2011; pp. 999–1006.
3. Chen, Y.; Li, W.; Chen, X.; Gool, L.V. Learning semantic segmentation from synthetic data: A geometrically guided input-output adaptation approach. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 1841–1850.
4. Chen, Y.; Li, W.; Sakaridis, C.; Dai, D.; Van Gool, L. Domain adaptive faster r-cnn for object detection in the wild. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 3339–3348.

5. Wilson, G.; Cook, D.J. A survey of unsupervised deep domain adaptation. *ACM Trans. Intell. Syst. Technol.* **2020**, *11*, 1–46. [CrossRef]
6. Wang, T.; Zhang, X.; Yuan, L.; Feng, J. Few-shot adaptive faster r-cnn. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 7173–7182.
7. Zhuang, C.; Han, X.; Huang, W.; Scott, M. ifan: Image-instance full alignment networks for adaptive object detection. In Proceedings of the AAAI Conference on Artificial Intelligence, New York, NY, USA, 7–12 February 2020; Volume 34; pp. 13122–13129.
8. Rezaeianaran, F.; Shetty, R.; Aljundi, R.; Reino, D.O.; Zhang, S.; Schiele, B. Seeking similarities over differences: Similarity-based domain alignment for adaptive object detection. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021; pp. 9204–9213.
9. Saito, K.; Ushiku, Y.; Harada, T.; Saenko, K. Strong-weak distribution alignment for adaptive object detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 6956–6965.
10. Xie, R.; Yu, F.; Wang, J.; Wang, Y.; Zhang, L. Multi-level domain adaptive learning for cross-domain detection. In Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops, Seoul, Republic of Korea, 27–28 October 2019.
11. Zhang, H.; Tian, Y.; Wang, K.; He, H.; Wang, F.Y. Synthetic-to-real domain adaptation for object instance segmentation. In Proceedings of the 2019 International Joint Conference on Neural Networks (IJCNN), Budapest, Hungary, 14–19 July 2019; pp. 1–7.
12. Chen, C.; Zheng, Z.; Ding, X.; Huang, Y.; Dou, Q. Harmonizing transferability and discriminability for adapting object detectors. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 8869–8878.
13. Zhu, X.; Pang, J.; Yang, C.; Shi, J.; Lin, D. Adapting object detectors via selective cross-domain alignment. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 687–696.
14. Wang, W.; Cao, Y.; Zhang, J.; He, F.; Zha, Z.J.; Wen, Y.; Tao, D. Exploring sequence feature alignment for domain adaptive detection transformers. In Proceedings of the 29th ACM International Conference on Multimedia, Chengdu, China, 20–24 October 2021; pp. 1730–1738.
15. Wang, K.; Pu, L.; Dong, W. Cross-domain Adaptive Object Detection Based on Refined Knowledge Transfer and Mined Guidance in Autonomous Vehicles. *IEEE Trans. Intell. Veh.* **2023**, *7*, 603–615. [CrossRef]
16. Cai, Q.; Pan, Y.; Ngo, C.W.; Tian, X.; Duan, L.; Yao, T. Exploring object relation in mean teacher for cross-domain detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 11457–11466.
17. Deng, J.; Li, W.; Chen, Y.; Duan, L. Unbiased mean teacher for cross-domain object detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Montreal, QC, Canada, 11–17 October 2021; pp. 4091–4101.
18. Li, Y.J.; Dai, X.; Ma, C.Y.; Liu, Y.C.; Chen, K.; Wu, B.; He, Z.; Kitani, K.; Vajda, P. Cross-domain adaptive teacher for object detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 7581–7590.
19. Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; Bengio, Y. Generative adversarial nets. In Proceedings of the Advances in Neural Information Processing Systems , Montreal, QC, Canada, 8–13 December 2014; Volume 27.
20. Hsu, H.K.; Yao, C.H.; Tsai, Y.H.; Hung, W.C.; Tseng, H.Y.; Singh, M.; Yang, M.H. Progressive domain adaptation for object detection. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, Snowmass Village, CO, USA, 1–5 March 2020; pp. 749–757.
21. Arruda, V.F.; Paixao, T.M.; Berriel, R.F.; De Souza, A.F.; Badue, C.; Sebe, N.; Oliveira-Santos, T. Cross-domain car detection using unsupervised image-to-image translation: From day to night. In Proceedings of the 2019 International Joint Conference on Neural Networks (IJCNN), Budapest, Hungary, 14–19 July 2019; pp. 1–8.
22. Oza, P.; Sindagi, V.A.; Sharmini, V.V.; Patel, V.M. Unsupervised domain adaptation of object detectors: A survey. *IEEE Trans. Pattern Anal. Mach. Intell.* **2023**, 1–24. [CrossRef]
23. Zhu, J.Y.; Park, T.; Isola, P.; Efros, A.A. Unpaired image-to-image translation using cycle-consistent adversarial networks. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 2223–2232.
24. Tarvainen, A.; Valpola, H. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In Proceedings of the Advances in Neural Information Processing Systems 30 (NIPS 2017), Long Beach, CA, USA, 4–9 December 2017; Volume 30.
25. Viola, P.; Jones, M. Rapid object detection using a boosted cascade of simple features. In Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR 2001, Kauai, HI, USA, 8–14 December 2001; Volume 1, p. I.
26. Viola, P.; Jones, M.J. Robust real-time face detection. *Int. J. Comput. Vis.* **2004**, *57*, 137–154. [CrossRef]
27. Dalal, N.; Triggs, B. Histograms of oriented gradients for human detection. In Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05), San Diego, CA, USA, 20–25 June 2005; Volume 1, pp. 886–893.
28. Felzenszwalb, P.F.; Girshick, R.B.; McAllester, D.; Ramanan, D. Object detection with discriminatively trained part-based models. *IEEE Trans. Pattern Anal. Mach. Intell.* **2009**, *32*, 1627–1645. [CrossRef] [PubMed]
29. Girshick, R.; Donahue, J.; Darrell, T.; Malik, J. Rich feature hierarchies for accurate object detection and semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 23–28 June 2014; pp. 580–587.

30. Girshick, R. Fast r-cnn. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 7–13 December 2015; pp. 1440–1448.
31. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster r-cnn: Towards real-time object detection with region proposal networks. In Proceedings of the Advances in Neural Information Processing Systems 28 (NIPS 2015), Montreal, QC, Canada, 7–12 December 2015; Volume 28.
32. He, K.; Gkioxari, G.; Dollár, P.; Girshick, R. Mask r-cnn. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 2961–2969.
33. Redmon, J.; Divvala, S.; Girshick, R.; Farhadi, A. You only look once: Unified, real-time object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 26 June–1 July 2016; pp. 779–788.
34. Redmon, J.; Farhadi, A. YOLO9000: Better, faster, stronger. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 7263–7271.
35. Redmon, J.; Farhadi, A. Yolov3: An incremental improvement. *arXiv* **2018**, arXiv:1804.02767.
36. Bochkovskiy, A.; Wang, C.Y.; Liao, H.Y.M. Yolov4: Optimal speed and accuracy of object detection. *arXiv* **2020**, arXiv:2004.10934.
37. Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S.; Fu, C.Y.; Berg, A.C. Ssd: Single shot multibox detector. In *Lecture Notes in Computer Science, Proceedings of the Computer Vision—ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, 11–14 October 2016*; Part I 14; Springer: Cham, Switzerland, 2016; pp. 21–37.
38. Lin, T.Y.; Goyal, P.; Girshick, R.; He, K.; Dollár, P. Focal loss for dense object detection. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 2980–2988.
39. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv* **2020**, arXiv:2010.11929.
40. Carion, N.; Massa, F.; Synnaeve, G.; Usunier, N.; Kirillov, A.; Zagoruyko, S. End-to-end object detection with transformers. In *Lecture Notes in Computer Science, Proceedings of the Computer Vision—ECCV 2020: 16th European Conference, Glasgow, UK, 23–28 August 2020*; Part I 16; Springer: Cham, Switzerland, 2020; pp. 213–229.
41. Zhu, X.; Su, W.; Lu, L.; Li, B.; Wang, X.; Dai, J. Deformable detr: Deformable transformers for end-to-end object detection. *arXiv* **2020**, arXiv:2010.04159.
42. Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; Guo, B. Swin transformer: Hierarchical vision transformer using shifted windows. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, QC, Canada, 11–17 October 2021; pp. 10012–10022.
43. Wang, W.; Xie, E.; Li, X.; Fan, D.P.; Song, K.; Liang, D.; Lu, T.; Luo, P.; Shao, L. Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, QC, Canada, 11–17 October 2021; pp. 568–578.
44. Lin, T.Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; Zitnick, C.L. Microsoft coco: Common objects in context. In *Lecture Notes in Computer Science, Proceedings of the Computer Vision—ECCV 2014: 13th European Conference, Zurich, Switzerland, 6–12 September 2014*; Part V 13; Springer: Cham, Switzerland, 2014; pp. 740–755.
45. Cao, S.; Joshi, D.; Gui, L.Y.; Wang, Y.X. Contrastive Mean Teacher for Domain Adaptive Object Detectors. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 18–22 June 2023; pp. 23839–23848.
46. Gong, K.; Li, S.; Li, S.; Zhang, R.; Liu, C.H.; Chen, Q. Improving Transferability for Domain Adaptive Detection Transformers. In Proceedings of the 30th ACM International Conference on Multimedia, Lisbon, Portugal, 10–13 October 2022; pp. 1543–1551.
47. Li, W.; Li, L.; Yang, H. Progressive cross-domain knowledge distillation for efficient unsupervised domain adaptive object detection. *Eng. Appl. Artif. Intell.* **2023**, *119*, 105774. [CrossRef]
48. Yu, F.; Wang, D.; Chen, Y.; Karianakis, N.; Shen, T.; Yu, P.; Lymberopoulos, D.; Lu, S.; Shi, W.; Chen, X. Sc-uda: Style and content gaps aware unsupervised domain adaptation for object detection. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, Waikoloa, HI, USA, 3–8 January 2022; pp. 382–391.
49. Liu, L.; Zhang, B.; Zhang, J.; Zhang, W.; Gan, Z.; Tian, G.; Zhu, W.; Wang, Y.; Wang, C. MixTeacher: Mining Promising Labels with Mixed Scale Teacher for Semi-Supervised Object Detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 18–22 June 2023; pp. 7370–7379.
50. Mattolin, G.; Zanella, L.; Ricci, E.; Wang, Y. ConfMix: Unsupervised Domain Adaptation for Object Detection via Confidence-based Mixing. In Proceedings of theIEEE/CVF Winter Conference on Applications of Computer Vision, Vancouver, BC, Canada, 18–22 June 2023; pp. 423–433.
51. Johnson-Roberson, M.; Barto, C.; Mehta, R.; Sridhar, S.N.; Rosaen, K.; Vasudevan, R. Driving in the matrix: Can virtual worlds replace human-generated annotations for real world tasks? *arXiv* **2016**, arXiv:1610.01983.
52. Yu, F.; Xian, W.; Chen, Y.; Liu, F.; Liao, M.; Madhavan, V.; Darrell, T. Bdd100k: A diverse driving video database with scalable annotation tooling. *arXiv* **2018**, arXiv:1805.04687.
53. Cordts, M.; Omran, M.; Ramos, S.; Rehfeld, T.; Enzweiler, M.; Benenson, R.; Franke, U.; Roth, S.; Schiele, B. The cityscapes dataset for semantic urban scene understanding. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 26 June–1 July 2016; pp. 3213–3223.
54. Geiger, A.; Lenz, P.; Stiller, C.; Urtasun, R. Vision meets robotics: The kitti dataset. *Int. J. Robot. Res.* **2013**, *32*, 1231–1237. [CrossRef]
55. Bottou, L. Stochastic gradient descent tricks. In *Neural Networks: Tricks of the Trade*, 2nd ed.; Springer: Berlin/Heidelberg, Germany, 2012; pp. 421–436.

56. Chen, K.; Wang, J.; Pang, J.; Cao, Y.; Xiong, Y.; Li, X.; Sun, S.; Feng, W.; Liu, Z.; Xu, J.; et al. MMDetection: Open mmlab detection toolbox and benchmark. *arXiv* **2019**, arXiv:1906.07155.
57. Paszke, A.; Gross, S.; Massa, F.; Lerer, A.; Bradbury, J.; Chanan, G.; Killeen, T.; Lin, Z.; Gimelshein, N.; Antiga, L.; et al. Pytorch: An imperative style, high-performance deep learning library. In Proceedings of the Advances in Neural Information Processing Systems 32 (NeurIPS 2019), Vancouver, BC, Canada, 8–14 December 2019; Volume 32.
58. Yoo, J.; Chung, I.; Kwak, N. Unsupervised domain adaptation for one-stage object detector using offsets to bounding box. In Proceedings of the European Conference on Computer Vision, ECCV 2022: Computer Vision—ECCV 2022, Tel Aviv, Israel, 23–27 October 2022; Springer: Cham, Switzerland, 2022; pp. 691–708.
59. Xu, C.D.; Zhao, X.R.; Jin, X.; Wei, X.S. Exploring categorical regularization for domain adaptive object detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 11724–11733.
60. Zhou, Q.; Gu, Q.; Pang, J.; Lu, X.; Ma, L. Self-adversarial disentangling for specific domain adaptation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2023**, *45*, 8954–8968. [CrossRef] [PubMed]
61. Jiang, J.; Chen, B.; Wang, J.; Long, M. Decoupled adaptation for cross-domain object detection. *arXiv* **2021**, arXiv:2110.02578.
62. Hsu, C.C.; Tsai, Y.H.; Lin, Y.Y.; Yang, M.H. Every pixel matters: Center-aware feature alignment for domain adaptive object detector. In *Lecture Notes in Computer Science, Proceedings of the Computer Vision—ECCV 2020: 16th European Conference, Glasgow, UK, 23–28 August 2020*; Part IX 16; Springer: Cham, Switzerland, 2020; pp. 733–748.
63. Li, W.; Liu, X.; Yuan, Y. Sigma: Semantic-complete graph matching for domain adaptive object detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 5291–5300.
64. Zhang, L.; Zhou, W.; Fan, H.; Luo, T.; Ling, H. Robust Domain Adaptive Object Detection with Unified Multi-Granularity Alignment. *arXiv* **2023**, arXiv:2301.00371.
65. Hnewa, M.; Radha, H. Integrated Multiscale Domain Adaptive YOLO. *IEEE Trans. Image Process.* **2023**, *32*, 1857–1867. [CrossRef] [PubMed]
66. Li, G.; Ji, Z.; Qu, X.; Zhou, R.; Cao, D. Cross-domain object detection for autonomous driving: A stepwise domain adaptative YOLO approach. *IEEE Trans. Intell. Veh.* **2022**, *7*, 603–615. [CrossRef]