applied sciences

MDPI

Article

Diverse Feature-Level Guidance Adjustments for Unsupervised Domain Adaptative Object Detection

Yuhe Zhu [1], Chang Liu [1], Yunfei Bai [1], Caiju Wang [1], Chengwei Wei [1], Zhenglin Li [2,3] and Yang Zhou [1,*]

[1] Research Institute of USV Engineering, School of Mechatronic Engineering and Automation, Shanghai University, Shanghai 200444, China; zyh15862623156@shu.edu.cn (Y.Z.); liuchang123@shu.edu.cn (C.L.); baiyunfei9527@shu.edu.cn (Y.B.); caijuwang@shu.edu.cn (C.W.); 21722259@shu.edu.cn (C.W.)
[2] Institute of Artifcial Intelligence, Shanghai University, Shanghai 200444, China; zhenglin_li@shu.edu.cn
[3] School of Future Technology, Shanghai University, Shanghai 200444, China
* Correspondence: saber_mio@shu.edu.cn

Abstract: **Unsupervised Domain Adaptative Object Detection** (UDAOD) aims to alleviate the gap between the source domain and the target domain. Previous methods sought to plainly align global and local features across domains but adapted numerous pooled features and overlooked contextual information, which caused incorrect perceptions of foreground information. To tackle these problems, we propose **Diverse Feature-level Guidance Adjustments** (DFGAs) for two-stage object detection frameworks, including **Pixel-wise Multi-scale Alignment** (PMA) and **Adaptative Threshold Confidence Adjustment** (ATCA). Specifically, PMA adapts features within diverse hierarchical levels to capture sufficient contextual information. Through a customized PMA loss, features from different stages of a network facilitate information interaction across domains. Training with this loss function contributes to the generation of more domain-agnostic features. To better recognize foreground and background samples, ATCA employs adaptative thresholds to divide the foreground and background samples. This strategy flexibly instructs the classifier to perceive the significance of box candidates. Comprehensive experiments are conducted on Cityscapes, Foggy Cityscapes, KITTI, and Sim10k datasets to further demonstrate the superior performance of our method compared to the baseline method.

Keywords: unsupervised domain adaptative object detection; feature distribution; feature alignment; foreground–background sample division

check for updates

Citation: Zhu, Y.; Liu, C.; Bai, Y.; Wang, C.; Wei, C.; Li, Z.; Zhou, Y. Diverse Feature-Level Guidance Adjustments for Unsupervised Domain Adaptative Object Detection. *Appl. Sci.* **2024**, *14*, 2844. https://doi.org/10.3390/app14072844

Academic Editor: Sungho Kim

Received: 28 December 2023
Revised: 12 March 2024
Accepted: 26 March 2024
Published: 28 March 2024

1. Introduction

Object detection aims to recognize and locate foreground objects within complex backgrounds. However, conventional detectors exhibit a noticeable performance decline when they are tested on real-world data. The root of this issue lies in the domain gap between training sets and real-world data. Conventional detectors are typically trained on labeled data, with researchers assuming that the training and validation sets follow an independent and identically distributed (i.i.d.) distribution. However, labeled samples are frequently absent when the detector is used in practical settings. Moreover, the distribution of training and validation sets may differ significantly in practical situations. This is because there are substantial differences in aspects such as illumination, shooting perspectives, and picture styles. **Unsupervised Domain Adaptative Object Detection** (UDAOD) has emerged as a solution to tackle this problem, as it aims to minimize the data distribution inconsistency across different domains. Researchers have introduced numerous UDAOD methods to effectively transfer models from labeled data to unlabeled data. These methods encompass abundant adaptation strategies, including data alignment and feature alignment strategies (image-level and instance-level [1], as shown in Figure 1) combined with an adversarial training [2] technique, among others.

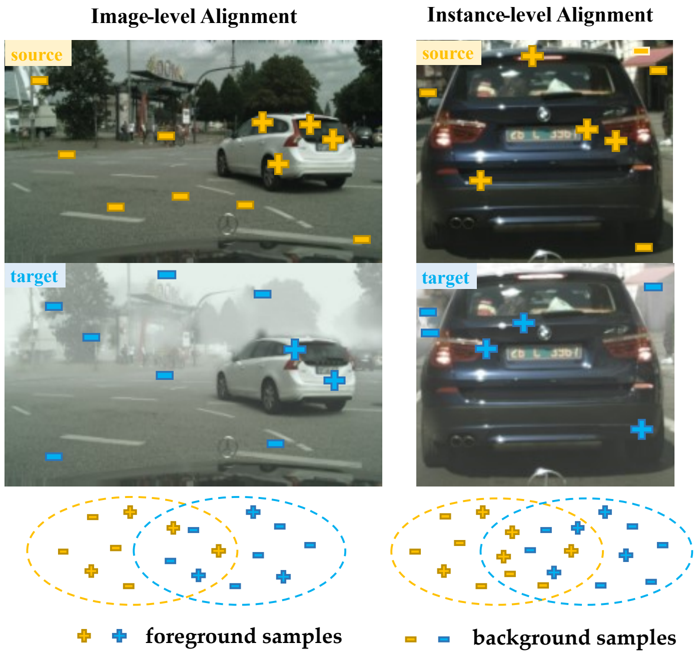


Figure 1. Both image-level and instance-level alignment approaches are commonly used in feature alignment methods. We use "+"and "−" to represent foreground and background samples, respectively. Additionally, the input requirements for these two alignment approaches differ. In image-level alignment, the entire set of image features is utilized to capture comprehensive information. In contrast, instance-level alignment concentrates on local information by leveraging proposal features as input.

These methods are typically based on one-stage or two-stage detectors. One-stage detectors [3–7] introduce image-level feature alignment with single stage detection frameworks. However, due to the absence of instance-level feature alignment, these methods suffer from incorrect awareness of foreground samples within complex backgrounds [5]. Conversely, two-stage object detection frameworks [1,8–16] utilize instance-level alignment to promote foreground perception through the Region Proposal Network (RPN) [17]. While an instance-level alignment strategy significantly improves object detection accuracy, it introduces certain issues. On the one hand, it hampers the discriminator's ability to distinguish features, since the pooled proposal features from the RPN commonly lose contextual and texture details. On the other hand, while using the RPN and an instance-level alignment strategy enhances domain invariant [12] proposal features, it also results in quantities of background candidate boxes.

To solve these problems, we introduce **Diverse Feature-level Guidance Adjustments** (DFGAs) for two-stage detection frameworks. To aid with the lack of context information [18] alignment, **Pixel-wise Multi-scale Alignment** (PMA) is proposed to adapt features on diverse hierarchical levels among different domains. Specifically, since global features contain comprehensive information, a novel similarity measurement loss function, termed PMA loss, is introduced to facilitate information exchange between the source and target domains. As for overly focusing on background candidates, the **Adaptative Threshold Confidence Adjustment** (ATCA) strategy is proposed as a category classifier. It encourages the model to focus on foreground information. Due to the necessity of foreground sample awareness, ATCA introduces adaptative thresholds to enhance the cross-domain alignment of foreground candidates. Specifically, it calculates the corresponding foreground–background sample division threshold based on the classifier's output. These samples are adaptively partitioned into allocation intervals (foreground, background, and ignored samples) depending on their corresponding thresholds. This approach enables the model to discern the characteristics of the objects and perceive the importance of the bounding box adaptatively.

We emphasize the contributions of this work as follows:

- A novel feature alignment strategy named Pixel-wise Multi-scale Alignment (PMA) is designed to minimize the contextual feature differences of the data distribution between the source domain and the target domain. Additionally, PMA instructs the backbone to generate more domain-agnostic feature representations.
- An effective sample division module named Adaptative Threshold Confidence Adjustment (ATCA) is proposed to guide the detector in better perceiving the importance of predictions for foreground and background samples. This approach effectively steers the model's perception towards the objects of interest as perceived.
- Extensive experiments were carried out on four benchmark datasets, Cityscapes, Foggy Cityscapes, KITTI, and Sim10k, to validate the effectiveness of the newly proposed modules. Our DFGAs improved by 1.3% mAP on weather adaptation and by 11.7% and 15.7% mAP on cross-camera adaptation, based on the same baseline method.

2. Related Works

2.1. Object Detection

Object detection [19] aims to identify and localize foreground objects from complex background information in a 2D image, without relying on depth approaches [20]. This technology enhances the efficiency and robustness of autonomous decision making within the realm of artificial intelligence. Grounded in deep learning, an object detection framework is categorized into one-stage detectors [21–23] and two-stage detectors [17,24–26]. One-stage detectors regress the location and category confidence of objects [27] directly. In contrast, two-stage detectors, equipped with the Region Proposal Network (RPN) [17], employ region proposals and refine bounding boxes continuously. Consequently, Faster R-CNN [17] is a well-known detection framework and serves as the foundation for numerous follow-up researches. Due to its flexibility, it is chosen as the baseline framework for many recent domain adaptative object detection methods.

2.2. Unsupervised Domain Adaptation for Object Detection

Compared to classic detectors, domain adaptation posits that there exists an inherent domain shift between training data and validation data. To mitigate the performance degradation caused by domain shift [1], researchers have proposed various methods based on Unsupervised Domain Adaptation. These methods include data alignment, feature alignment, and adversarial training [2], as well as other strategies like disentangled representation learning [28] and style-transferred methods [29], and so on.

Regarding data distribution differences, Yang et al. [30] and Liu et al. [31] incorporated the exchange of features' low-frequency spectral information in the frequency domain between different domains. Furthermore, Hsu et al. [32] proposed the intermediate domain, which generated synthetic data by Cycle GAN [33] to mimic target domain data. Nevertheless, the effectiveness of domain adaptation through data-level alignment is limited. Researchers then shifted their attention to feature adaptation through feature alignment. Methods like DA Faster R-CNN [1] pioneered the introduction of both image-level and instance-level feature alignment strategies, incorporating consistency regularization based on the prediction of the two classifiers. Xu et al. [8] measured the prediction of image-level and instance-level classifiers. The matching results are produced and used as standards to separate samples of the foreground objects. Zhou et al. [9] achieved feature alignment through multi-granularity level alignment strategies, such as pixel-level, instance-level, and category-level strategies.

Since applying local regions' information [14–16,34–36] with instance-level alignment, these recent works have achieved significant improvements in detection performance. However, domain adaptation requires more than the alignment of features from backbone or RPN. It is fundamental to take contextual connections between features into account.

Our PMA loss compensates for the insufficient information interaction between the source domain and the target domain in context. Furthermore, enhancing the perception and alignment of foreground information is crucial and cannot be ignored. Although some methods [1,8,9] adopt category consistency regularization to enhance the alignment of category information, they also overfit substantial background information. To selectively focus on foreground information, ATCA is proposed for adaptively partitioning samples as foreground samples, background samples, and ignored samples. Consequently, based on the Faster R-CNN [17] framework, a pixel-wise contextual alignment strategy and an adaptative sample division strategy are presented, as detailed in the next section.

3. Methods

This section illustrates the overall algorithm framework, as well as two proposed strategies aiming to enhance the accuracy of **Unsupervised Domain Adaptative for Object Detection** (UDAOD). Section 3.1 introduces the framework overview of the DFGAs method. In Section 3.2 and Section 3.3, Pixel-wise Multi-scale Alignment and Adaptative Threshold Confidence Adjustment are detailed, respectively.

3.1. Framework Overview

Our DFGAs follow the two-stage detector Faster R-CNN [17], regarding its foundational architecture. The general framework overview is presented in Figure 2. We employ an unsupervised domain adaptation baseline for cross-domain object detection. During the training phase, labeled source domain data and unlabeled target domain data are jointly used for feature extraction and distribution discrimination through adversarial training. This encourages the feature generator to generate more high-quality domain-invariant features. It also contributes to mitigating the impact of specific domain data styles when detecting objects. To guide this process, two plug-and-play modules are proposed: the pixel-wise multi-scale alignment module and the adaptative threshold confidence adjustment module.

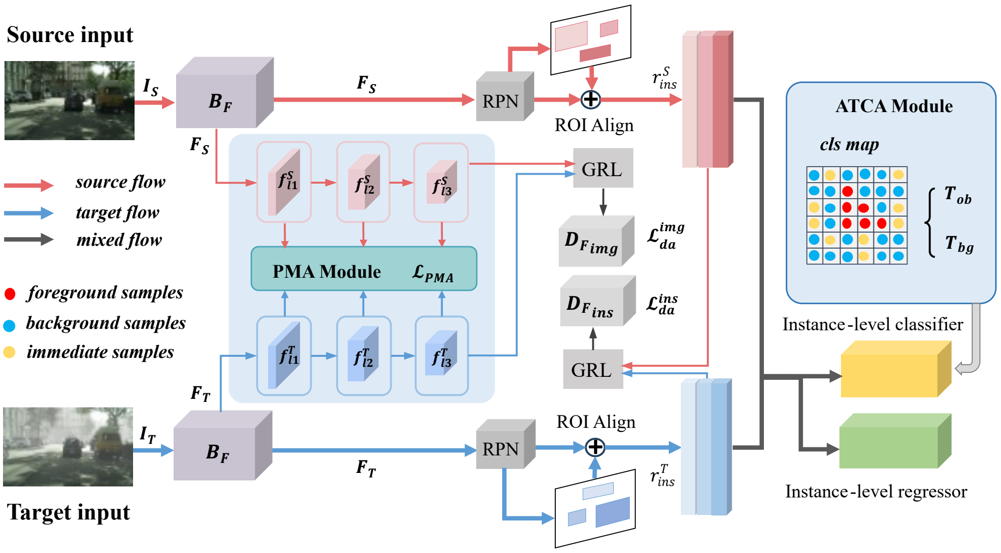


Figure 2. An overview of our Diverse Feature-level Guidance Adjustment. Training batch includes labeled data I_S and unlabeled data I_T. The two proposed plug-and-play components are PMA module and ATCA module, as illustrated. PMA module is performed on image-level features at different hierarchical layers. The predictions obtained from the instance-level classifiers are fed into ATCA module for sample division. Feature flows F_S and F_T facilitate gradient propagation via adversarial training and GRL (Gradient Reversal Layer) between the feature generator B_F and domain discriminator D_F. The instance-level representations r^S_{ins} and r^T_{ins} generated by RPN and ROI Align are fed into the detection head (instance-level classifier and classifier).

3.2. Pixel-Wise Multi-Scale Alignment

In order to enhance feature adaptation, Pixel-wise Multi-scale Alignment (PMA) is proposed to capture sufficient contextual information. Following the regular training methods of domain adaptation, we resort to adversarial training [2] techniques. To narrow the data distribution gap across different domains, the PMA module is applied to features from different backbone stages. The implementation process of PMA is as shown in Figure 3.

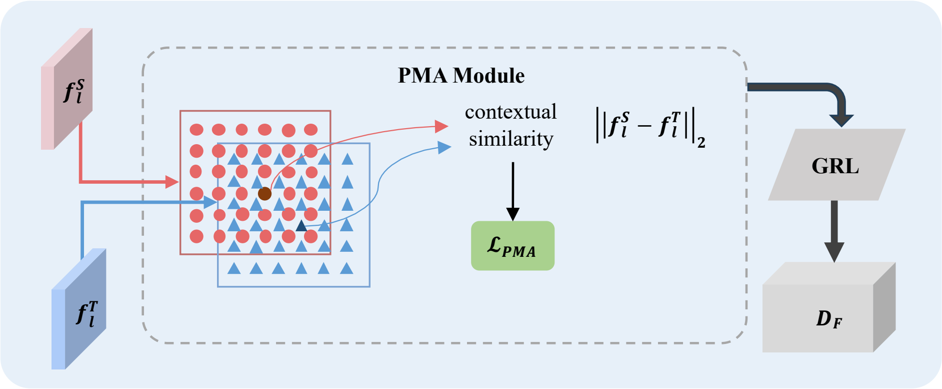


Figure 3. Proposed pixel-level multi-scale alignment. Given the output features of each stage of the network across two domains, PMA calculates the similarity between source features and target features for each batch. PMA loss is then performed on the batched features to align the two domain data distributions.

During the training phase, multi-scale features from the source domain and the target domain are calculated, using the PMA module pixel-wisely. To measure the similarity between the two domains within the same batch, Sim_{pixel} is introduced, which quantifies the degree of similarity between them, as can be seen in Figure 3. The similarity computation of cross-domain features is as follows:

$$\mathrm{Sim}_{pixel}\left(f_l^S, f_l^T\right) = \left\|f_l^S - f_l^T\right\|_2, \tag{1}$$

where f_l^S and f_l^T denote the features from the source domain and the target domain, respectively.

Inspired by TIA [16], the VGG16 [37] network is divided into three stages. Subsequently, PMA loss, represented by $\mathcal{L}_{PMA}$, is constructed to bring the feature distributions of various domains closer together, as shown in Equation (2). i denotes the ith stage of the network, and is set to 1, 2, and 3. X and Y are the sample feature sets from the three stages of the backbone network. a and b are current features from X and Y, respectively, used for calculating the PMA loss. The contextual similarity is taken as an exponential factor [38] in calculating the similarity learning loss. The loss function formulated by the PMA module is as follows:

$$\mathcal{L}_{PMA}\left(f_l^S, f_l^T\right) = -\log \frac{\sum_{a \in X_i} e^{\mathrm{Sim}_{pixel}\left(f_{l_i}^S, f_{l_i}^T\right)}}{\sum_{a \in X_i} e^{\mathrm{Sim}_{pixel}\left(f_{l_i}^S, f_{l_i}^T\right)} + \sum_{b \in Y_i} e^{\mathrm{Sim}_{pixel}\left(f_{l_i}^S, f_{l_i}^T\right)}}, \quad i = 1, 2, 3 \tag{2}$$

Since there are inconsistencies in the feature representations at different stages, we choose to employ different weights to assess the importance of features. Three parameters are commonly set, with $\alpha = 0.2$, $\beta = 0.3$, and $\gamma = 0.5$. The total PMA loss is formulated as follows:

$$\mathcal{L}_{PMA} = \alpha \mathcal{L}_{PMA}^1 + \beta \mathcal{L}_{PMA}^2 + \gamma \mathcal{L}_{PMA}^3 \tag{3}$$

The features generated by the feature generator B_F are $B \times C \times W \times H$ tensors. B denotes the batch size of the input in the training phase. C, W, and H are the channel, width, and height of the input images, respectively. When calculating the similarity of features based on the PMA loss, the computation is performed along the channel dimension. However, due to the excessive computational burden imposed by channel-wise calculations, cross-channel calculations become necessary. At this point, it is validated that a stride of 16 achieves optimal performance, and this will be detailed in Section 4.4.

Our PMA loss enhances the feature contextual alignment across domains and provides guidance for the domain discriminator. The overall loss of the method is formulated as follows:

$$\mathcal{L} = \mathcal{L}_{\text{det}} + \mathcal{L}_{da} + \mathcal{L}_{PMA} \,, \tag{4}$$

where $\mathcal{L}_{\text{det}}$ represents the training loss of the baseline Faster R-CNN, and $\mathcal{L}_{da}$ denotes the adversarial training loss provided by the domain discriminator.

3.3. Adaptative Threshold Confidence Adjustment

The ultimate goal of UDAOD remains to achieve object detection in unlabeled data. The performance of the detector relies heavily on the classifier. Therefore, it is necessary to achieve foreground alignment between domains. For sample classification, cross-entropy loss is employed. The classification loss is calculated as follows:

$$\mathcal{L}_{\text{class}}^{ins}\left(X_j, \widehat{X_j}\right) = -\frac{1}{N}\sum_{j=1}^{N} X_j \times \log\left(\widehat{X_j}\right) , \tag{5}$$

where X_j denotes the label value of the jth sample X and $\widehat{X_j}$ means the values predicted by the classifier correspondingly. N is the current total number of samples.

In order to enhance the cross-domain alignment of foreground sample information, an adaptative threshold adjustment strategy, ATCA, is proposed to diminish the confusion caused by background samples. Based on the confidence of the current batch samples, the mean value and variance are calculated in Equation (6). T_{ob} and T_{bg} are employed to partition positive and negative samples when assigning category labels, as detailed in Equation (7), as follows:

$$\mu_i = \frac{1}{K}\sum_{K}^{i=1} p_i, \quad v_i = \frac{1}{K}\sum_{K}^{i=1}\sqrt{(p_i - \mu)^2} \,, \tag{6}$$

$$T_{ob} = \mu_i + v_i, \quad T_{bg} = \mu_i - v_i \,, \tag{7}$$

where μ_i and v_i represent the mean value and variance of the confidence for the ith sample, respectively. p_i represents the prediction results and K is the total numbers of the samples.

This approach allows for a more reasonable division of highly confident positive samples and low-confidence negative samples, with the sample division defined as follows:

$$X_j \begin{cases} \text{foreground sample} \,, \widehat{P_{\text{out}}} > T_{ob} \\ \text{background sample} \,, \widehat{P_{\text{out}}} < T_{bg} \\ \text{ignored sample} \,, T_{bg} < \widehat{P_{out}} < T_{ob} \end{cases} , \tag{8}$$

where X_j is the jth sample and $\widehat{P_{\text{out}}}$ denotes the sample confidence predicted by the classifier for each category within every bounding box.

Classifier inputs comprise both image-level and instance-level features, enabling predictions at corresponding granularities. Global feature predictions provide a general foreground assessment, whereas instance-level classifiers using features refined by RPN [17] and ROI Align enhance the accuracy of foreground bounding boxes. We introduced object consistency regularization to harmonize the prediction following image-level alignment and instance-level alignment. Therefore, this deviation from coarse to fine perception with image-level to instance-level is measured by Equation (9). At this point, the importance of foreground objects is reflected in the Weight$_{ins}$, as follows:

$$\mathrm{m}(\cdot) = e^{\widehat{P_{ins}} - \widehat{P_{img}}}\ , \tag{9}$$

$$\text{Weight}\ _{ins} = \widehat{P_{ins}} \times \mathrm{m}(\cdot)\ , \tag{10}$$

where $\widehat{P_{ins}}$ and $\widehat{P_{img}}$ denote the predictions output by the instance-level and image-level classifiers, respectively. $\mathrm{m}(\cdot)$ is the difference between the outputs of the two classifiers. Weight$_{ins}$ is designed to describe the reliability of the predicted foreground class results within the bounding boxes, by weighting the entropy using the prediction of distinct classes to focus on the proper category. Typically, a standard entropy loss is used to classify foreground objects, denoted as $\mathcal{L}_{\text{class}}^{G-L}$, and the ultimate global–local consistency loss is formulated as:

$$\mathcal{L}_{\text{class}}^{G-L} = \text{Weight}\ _{ins} \times \mathcal{L}_{\text{class}}^{ins} \tag{11}$$

4. Experiments

In this section, the evaluation of the proposed DFGAs framework is detailed. Section 4.1 introduces the implementation details of our experiments. Section 4.2 briefly illustrates the benchmark datasets used for experimental verification. In Section 4.3, the proposed method is evaluated by comparing it with seven UDAOD methods, and the results are shown with mAP(%). A detailed ablation study is designed to analyze the effectiveness of the proposed modules in Section 4.4. Finally, Section 4.5 presents the analysis, together with visualizations of the detection results.

4.1. Implementation Details

Following the recent methods in UDAOD, our experimental base detection model adopted the Faster R-CNN [17] framework, with VGG-16 [37] as the backbone. In all experiments, the input images were resized to make the shorter side equal to 600 pixels and longer side less or equal to 1200 pixels. For experiments on *weather adaptation* and *synthetic-to-real adaptation*, the training iterations were set to 100k, while for the experiments on *cross-camera adaptation*, 70k iterations were trained in total. The initial learning rate was set to 0.001 and decayed by a factor of 10 every 50k iterations. We set the batch size as two and saved the model weights every 10k iterations during the training phase. As for the loss function, the parameters α, β, and γ were set to 0.2, 0.3, and 0.5, respectively.

4.2. Datasets

Our experiments set up three main adaptative scenarios, including *weather adaptation*, *cross-camera adaptation*, and *synthetic-to-real adaptation*, corresponding to four benchmark datasets: Cityscapes [39], Foggy Cityscapes [38], KITTI [40], and Sim10k [41].

Weather Adaptation. The Cityscapes [39] and Foggy Cityscapes [38] datasets were employed to study domain shift caused by weather conditions. The Cityscapes dataset is a dataset of urban scenes captured under dry weather conditions. The Foggy Cityscapes dataset was created by adding artificial fog to Cityscapes. For our detection tasks, we focused on eight specific categories: *car*, *train*, *motorcycle*, *bus*, *bicycle*, *cycle*, *person*, and *truck*, to evaluate the accuracy of the proposed algorithm. A total of 2965 images from the Cityscapes and Foggy Cityscapes datasets were used as the source and target domain input data, respectively. An additional 492 images were selected from the Foggy Cityscapes dataset to create the validation set.

Cross-camera Adaptation. The Cityscapes [39] and KITTI [40] datasets were utilized to evaluate the model's cross-camera adaptation capability. These two datasets have different camera configurations. The KITTI dataset, which includes urban, rural, and highway images in real scenarios, comprises 7481 images. The *car* category was primarily analyzed to assess our method’s superiority. The training set input comprised 2965 Cityscapes images and all KITTI images. A total of 500 images were selected from different training domain datasets as the validation set.

Synthetic-to-real Adaptation. To assess the adaptation ability from real to synthetic images, the transition from Cityscapes [39] to Sim10k [41] was utilized for evaluation. Similarly, *car* category bounding boxes were introduced as the primary objects in our experiments. A total of 2965 images from the Cityscapes dataset and all images from the Sim10k dataset were used as the target and source domain input, respectively. To create the validation set, 500 images were selected from different training sets of the Cityscapes dataset.

4.3. Comparative Experiments

We have conducted comparative experiments with several UDAOD methods in recent years, including Faster R-CNN [17], DA Faster R-CNN [1], ATF [12], SAP [35], MeGA [15], SWDA [14], TIA [16], and MAF [42]. These methods are grounded in the Faster R-CNN [17] framework and the VGG16 [37] network. Our research focuses on addressing the deficiencies associated with different feature alignment methods, including global and local feature alignment. We chose the above methods to conduct our experiment.

4.3.1. Weather Adaptation

According to Table 1, our DFGAs achieved the highest mAP of 43.6%, marking an approximate 1.3% mAP improvement over TIA. This result indicates that our approach outperformed the previous methods in UDAOD regarding detection accuracy. It demonstrates the significance of aligning the feature context information and enhancing foreground information cognition on weather adaptation.

Table 1. Experimental results (%) of DFGAs method on the Cityscapes → Foggy Cityscapes compared with other methods.

Method	**Bus**	**Bicycle**	**Car**	**Mcycle**	**Person**	**Rider**	**Train**	**Truck**	**mAP**
Faster R-CNN	25.5	30.3	34.2	19.0	25.1	33.4	9.1	12.1	23.7
DA Faster R-CNN	35.3	27.1	40.5	20.0	25.0	31.0	20.2	22.1	27.6
SDWA	36.2	35.3	43.5	30.0	29.9	42.3	32.6	24.5	34.3
ATF	43.3	38.8	50.0	33.4	34.6	47.0	38.7	23.7	38.7
SAP	46.8	40.7	**59.8**	30.4	**40.8**	46.7	37.5	24.3	40.9
MeGA	49.2	39.0	52.4	34.5	37.7	**49.0**	46.9	25.4	42.8
TIA	52.1	38.1	49.7	37.7	34.8	46.3	**48.6**	31.1	42.3
Ours	**53.6**	**41.6**	53.0	**40.1**	36.8	47.9	41.3	**34.3**	**43.6**

4.3.2. Cross-Camera Adaptation

To validate the effectiveness of our method, experiments were conducted, not only for weather differences in paired datasets, but also for adapting to camera distinctions, using the Cityscapes [39] and KITTI [40] datasets, as presented in Table 2. We utilized these datasets as the inputs for the source and target domains, respectively. According to Table 2, DFGAs achieved 59.5% mAP and 87.6% mAP in detecting cars. In comparison with TIA, DFGAs achieved improvements of approximately 15.5% mAP in KITTI → City and 11.7% mAP in City → KITTI. It is evident that our method exhibited remarkable performances when mitigating the domain shift arising from camera differences.

Table 2. Experimental results (%) of DFGAs on KITTI → Cityscapes and Cityscapes → KITTI datasets compared with other methods.

Method	KITTI → City (mAP)	City → KITTI (mAP)
Faster R-CNN	30.2	53.5
DA Faster R-CNN	38.5	64.1
SWDA	37.9	71.0
SAP	43.4	75.2
ATF	42.1	73.5
MeGA	43.0	75.5
TIA	44.0	75.9
DFGAs	**59.5**	**87.6**

4.3.3. Synthetic-to-Real Adaptation

Regarding the domain adaptation between synthetic data and real-world data, we conducted experiments on Sim10K [41] and Cityscapes [39] datasets. Table 3 shows that our DFGAs surpassed all other methods, with 41.6% mAP. This illustrates that our method bridges the differences between synthetic data and real data, to some extent.

Table 3. Experimental results (%) of DFGAs on Sim10K → Cityscapes datasets compared with other methods.

Method	Sim10K → City (mAP)
DA Faster R-CNN	34.3
TIA	39.6
SWDA	40.1
MAF	41.1
DFGAs	**41.6**

4.4. Ablation Study

The ablation experiment was designed to investigate the unique facilitative effects of the DFGAs method in UDAOD. This study focused on evaluating module effectiveness on the Cityscapes and Foggy Cityscapes datasets. Each module was independently integrated into the codebase and its effectiveness was empirically validated.

We employed the UDAOD model TIA [16], based on the Faster R-CNN framework, for our first ablation experiment. The experimental results are presented in Table 4. Model A was the baseline model (TIA), which had 42.3% mAP. When we incorporated a PMA module, mAP was improved by 0.9%. Model C, with the ATCA module, contributed an additional 0.7% mAP value. Finally, Model D was equipped with the PMA and ATCA modules and achieved 43.6% mAP. These improvements confirm the role of PMA loss in aligning feature distributions across domains, while our ATCA module effectively enhanced the classifier's focus on foreground information.

Table 4. Results (%) of ablation experiment on the baseline TIA.

Model	PMA	ATCA	mAP
A	–	–	42.3
B	✓	–	43.2
C	–	✓	43.0
D	✓	✓	43.6

To demonstrate the generality of our modules, we integrated them into another model, CRDA [8], as viewed in Table 5. Model E was the baseline model of CRDA, which achieved

29.9% mAP. Obviously, our PMA improved the mAP by 2.2% in Model F and ATCA improved it by 1.9% mAP in Model G, respectively. In Model H, the modules together yielded 2.5% mAP improvements. These substantial mAP improvements reinforced the effectiveness of our method, which offered plug-and-play convenience and was effective in cross-domain object detection.

Table 5. Results (%) of ablation experiment on the baseline CRDA.

Model	PMA	ATCA	mAP
E	–	–	29.9
F	✓	–	32.1
G	–	✓	31.8
H	✓	✓	32.4

To investigate the influence of the stride of cross-channel calculation in the PMA module on domain adaptative object detection accuracy, we conducted a series of experimental training and validations on the baseline TIA [16]. Experiments with strides of 1, 2, 8, 16, and 32 were conducted. The findings, depicted in Figure 4, revealed that the PMA module reached peak performance with a stride of 16.

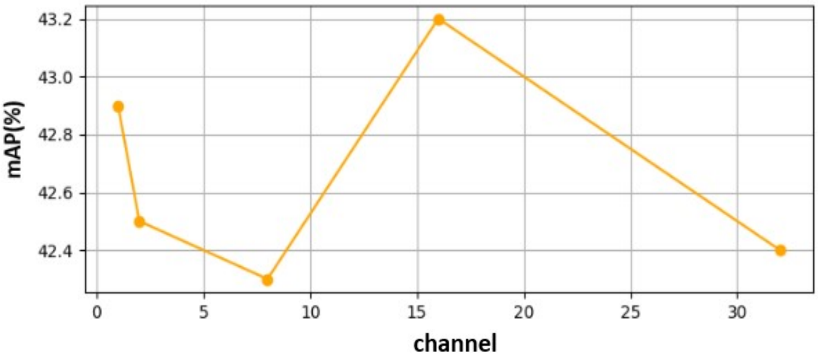


Figure 4. mAP(%) varies with the stride of channels within the PMA module.

For our ATCA strategy, we assessed the effects of fixed and adaptative thresholds on sample division through comparative experiments. We trained the model with thresholds fixed at 0.5, 0.7, and 1 based on the baseline CRDA [8], and the results are presented in Figure 5. Ultimately, our adaptative threshold sample division strategy, ATCA, outperformed the fixed threshold approach clearly.

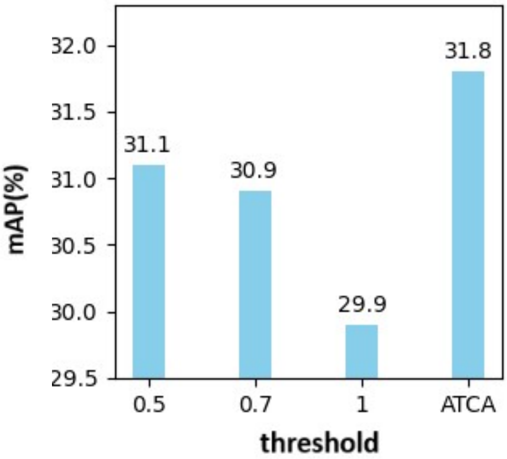
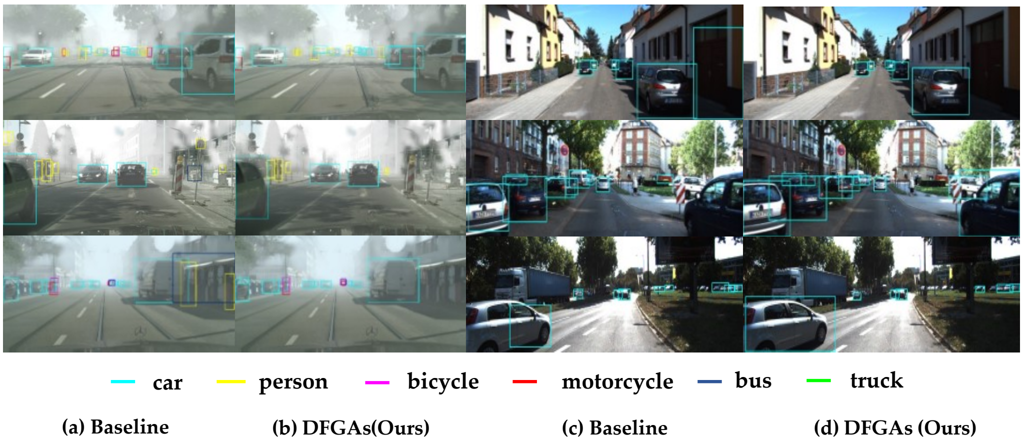


Figure 5. mAP(%) varies with threshold within the ATCA module.

4.5. Analysis

We compared the visualization results for the *weather adaptation* and *cross-camera adaptation* experiments. Despite the existence of an unavoidable gap between the source and target domains, PMA loss was used to narrow the divergence between their features using contextual information. Moreover, the ATCA strategy prevents many background candidate boxes to some extent.

Figure 6 displays the prediction results of both the baseline method and our DFGAs method in the target domain for the *weather adaptation* and *cross-camera adaptation* experiments. The *weather adaptation* experiment results, shown in Figure 6a,b, illustrate that our DFGAs method reduced some false detection on the background compared to the baseline. As for the *cross-camera adaptation* results presented in Figure 6c,d, there were also noticeable corrections of detection failure and detection fault. Under identical threshold conditions, our method reduced false detection, notably by preventing the generation of background candidate boxes. It is evident that DFGAs yielded superior bounding boxes compared to the baseline method. Specifically, DFGAs demonstrated a more accurate perception of both foreground samples and background samples in these two experiment sets. DFGAs not only enhanced foreground alignment across varied domains but also minimized the emergence of negative samples.



(a) Baseline (b) DFGAs(Ours) (c) Baseline (d) DFGAs (Ours)

Figure 6. Comparative experiments results' visualization on weather adaptation (Foggy Cityscapes) and cross-camera adaptation (KITTI) between the baseline and DFGAs method. (**a**,**b**) represent the results of the baseline and DFGAs method for weather adaptation, respectively; (**c**,**d**) are the results of the baseline and DFGAs method for cross-camera adaptation, respectively.

5. Discussion

In this paper, we propose a method called DFGAs for achieving cross-domain object detection in unsupervised domain adaptation scenarios. Previous approaches based on the Faster R-CNN framework often utilize feature alignment strategies to mitigate cross-domain data distribution discrepancies. Our research primarily focuses on addressing the deficiencies in some feature alignment methods. With image-level feature alignment, maximizing cross-domain information interaction while aligning global features become necessary. PMA loss is proposed to measure feature similarity across multiple scales. Another key focus of this paper is to mitigate the influence of background information on feature alignment in instance-level features. The ATCA module adaptively adjusts classification thresholds to flexibly categorize sample information. Our proposed feature alignment method simplifies and enhances the alignment of source and target domain feature distributions, with a more direct and explicit approach.

However, there is some potential future work worth exploring. Firstly, we plan to investigate additional feature-level alignment techniques for further reducing cross-domain data distribution disparities to improve unsupervised domain adaptative object detection accuracy. Secondly, while our experiments were conducted on four sets of datasets, we are preparing to further broaden the applicability of our method on diverse datasets. Thirdly, we aim to assess our method's integration in real-world scenarios, targeting on-site deployment for effective environmental object monitoring.

6. Conclusions

This paper introduces DFGAs, a novel approach to enhance object detection in unlabeled target domain datasets. One of our contributions lies in proposing the PMA loss to enhance the interaction of global features across domains, thereby reducing the distribution gap between different domains. Additionally, we propose the use of adaptative thresholds to classify foreground and background samples more flexibly for RPN features. Our ATCA module improves the classifier's accuracy in perceiving foreground information, aiding the model in focusing on foreground details. We conducted extensive experiments and ablation studies to validate the effectiveness of each proposed component. In future research, we will continue to explore domain shift solutions in object detection, focusing on cross-domain feature alignment among other approaches.

Author Contributions: Conceptualization, Y.Z. (Yuhe Zhu) and Y.Z. (Yang Zhou); methodology, Y.Z. (Yuhe Zhu) and C.L.; software, Y.Z. (Yuhe Zhu) and Y.B.; validation, Y.Z. (Yuhe Zhu) and C.W. (Caiju Wang); formal analysis, Y.Z. (Yuhe Zhu) and C.W. (Chengwei Wei); investigation, Y.Z. (Yuhe Zhu) and Y.B.; resources, Y.Z. (Yuhe Zhu); data curation, Y.Z. (Yuhe Zhu) and C.W. (Caiju Wang); writing—original draft preparation, Y.Z. (Yuhe Zhu) and C.L.; writing—review and editing, Y.Z. (Yuhe Zhu) and Z.L.; visualization, Y.Z. (Yuhe Zhu) and C.W. (Chengwei Wei); supervision, Y.Z. (Yang Zhou); project administration, Z.L.; funding acquisition, Y.Z. (Yang Zhou). All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Young Scientists Fund of the National Natural Science Foundation of China (Grant No. 62203289), the Shanghai Sailing Program (Grant No. 22YF1413800), and the National Natural Science Foundation of China (Grant No. 52371371).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The original contributions presented in the study are included in the article, further inquiries can be directed to the corresponding author.

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

References

1. 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.
2. Ganin, Y.; Ustinova, E.; Ajakan, H.; Germain, P.; Larochelle, H.; Laviolette, F.; March, M.; Lempitsky, V. Domain-adversarial training of neural networks. *J. Mach. Learn. Res.* **2016**, *17*, 1–35.
3. Chen, C.; Zheng, Z.; Huang, Y.; Ding, X.; Yu, Y. I3net: Implicit instance-invariant network for adapting one-stage object detectors. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 12576–12585.
4. Zhang, S.; Tuo, H.; Hu, J.; Jing, Z. Domain adaptive YOLO for one-stage cross-domain detection. In Proceedings of the Asian Conference on Machine Learning. PMLR, London, UK, 8–11 November 2021; pp. 785–797.
5. 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 Proceedings of the Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, 23–28 August 2020; Proceedings, Part IX 16; Springer: Berlin/Heidelberg, Germany, 2020; pp. 733–748.
6. Hnewa, M.; Radha, H. Multiscale domain adaptive yolo for cross-domain object detection. In Proceedings of the 2021 IEEE International Conference on Image Processing (ICIP), Anchorage, AL, USA, 19–22 September 2021; IEEE: Piscataway, NJ, USA, 2021; pp. 3323–3327.
7. Liu, W.; Ren, G.; Yu, R.; Guo, S.; Zhu, J.; Zhang, L. Image-adaptive YOLO for object detection in adverse weather conditions. In Proceedings of the AAAI Conference on Artificial Intelligence, Vancouver, BC, Canada, 20–27 February 2022; Volume 36, pp. 1792–1800.
8. 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.

9. Zhou, W.; Du, D.; Zhang, L.; Luo, T.; Wu, Y. Multi-granularity alignment domain adaptation 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. 9581–9590.
10. Sindagi, V.A.; Oza, P.; Yasarla, R.; Patel, V.M. Prior-based domain adaptive object detection for hazy and rainy conditions. In Proceedings of the Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, 23–28 August 2020; Proceedings, Part XIV 16; Springer: Berlin/Heidelberg, Germany, 2020; pp. 763–780.
11. Jiang, J.; Chen, B.; Wang, J.; Long, M. Decoupled adaptation for cross-domain object detection. *arXiv* **2021**, arXiv:2110.02578.
12. He, Z.; Zhang, L. Domain adaptive object detection via asymmetric tri-way faster-rcnn. In Proceedings of the Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, 23–28 August 2020; Proceedings, Part XXIV 16; Springer: Berlin/Heidelberg, Germany, 2020; pp. 309–324.
13. Li, S.; Huang, J.; Hua, X.S.; Zhang, L. Category dictionary guided unsupervised domain adaptation for object detection. In Proceedings of the AAAI Conference on Artificial Intelligence, Philadelphia, PA, USA, 2–9 February 2021; Volume 35, pp. 1949–1957.
14. 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.
15. Vs, V.; Gupta, V.; Oza, P.; Sindagi, V.A.; Patel, V.M. Mega-cda: Memory guided attention for category-aware unsupervised domain adaptive object detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Virtual, 19–25 June 2021; pp. 4516–4526.
16. Zhao, L.; Wang, L. Task-specific inconsistency alignment 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. 14217–14226.
17. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster r-cnn: Towards real-time object detection with region proposal networks. *Adv. Neural Inf. Process. Syst.* **2015**, *28*, 1–9. [CrossRef] [PubMed]
18. Xie, S.; Liu, C.; Gao, J.; Li, X.; Luo, J.; Fan, B.; Chen, J.; Pu, H.; Peng, Y. Diverse receptive field network with context aggregation for fast object detection. *J. Vis. Commun. Image Represent.* **2020**, *70*, 102770. [CrossRef]
19. Zou, Z.; Chen, K.; Shi, Z.; Guo, Y.; Ye, J. Object detection in 20 years: A survey. *Proc. IEEE* **2023**, *111*, 257–276. [CrossRef]
20. Bulbul, M.F.; Tabussum, S.; Ali, H.; Zheng, W.; Lee, M.Y.; Ullah, A. Exploring 3D human action recognition using STACOG on multi-view depth motion maps sequences. *Sensors* **2021**, *21*, 3642. [CrossRef] [PubMed]
21. 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, 27–30 June 2016; pp. 779–788.
22. Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S.; Fu, C.Y.; Berg, A.C. Ssd: Single shot multibox detector. In Proceedings of the Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, 11–14 October 2016; Proceedings, Part I 14; Springer: Berlin/Heidelberg, Germany, 2016; pp. 21–37.
23. Tian, Z.; Shen, C.; Chen, H.; He, T. Fcos: Fully convolutional one-stage object detection. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 9627–9636.
24. Girshick, R. Fast r-cnn. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 7–13 December 2015; pp. 1440–1448.
25. 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.
26. Cai, Z.; Vasconcelos, N. Cascade r-cnn: Delving into high quality object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 6154–6162.
27. Liu, C.; Xie, S.; Li, X.; Gao, J.; Xiao, W.; Fan, B.; Peng, Y. Mitigate the classification ambiguity via localization-classification sequence in object detection. *Pattern Recognit.* **2023**, *138*, 109418. [CrossRef]
28. 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]
29. Menke, M.; Wenzel, T.; Schwung, A. AWADA: Attention-Weighted Adversarial Domain Adaptation for Object Detection. *arXiv* **2022**, arXiv:2208.14662.
30. Yang, Y.; Soatto, S. Fda: Fourier domain adaptation for semantic segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Washington, DC, USA, 14–19 June 2020; pp. 4085–4095.
31. Liu, R.; Han, Y.; Wang, Y.; Tian, Q. Frequency Spectrum Augmentation Consistency for Domain Adaptive Object Detection. *arXiv* **2021**, arXiv:2112.08605.
32. 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, Village, CO, USA, 1–5 March 2020; pp. 749–757.
33. 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.
34. Shen, Z.; Maheshwari, H.; Yao, W.; Savvides, M. Scl: Towards accurate domain adaptive object detection via gradient detach based stacked complementary losses. *arXiv* **2019**, arXiv:1911.02559.
35. Li, C.; Du, D.; Zhang, L.; Wen, L.; Luo, T.; Wu, Y.; Zhu, P. Spatial attention pyramid network for unsupervised domain adaptation. In Proceedings of the Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, 23–28 August 2020; Proceedings, Part XIII 16; Springer: Berlin/Heidelberg, Germany, 2020; pp. 481–497.

36. Sharma, A.; Kalluri, T.; Chandraker, M. Instance level affinity-based transfer for unsupervised domain adaptation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Virtual, 19–25 June 2021; pp. 5361–5371.
37. Simonyan, K.; Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv* **2014**, arXiv:1409.1556.
38. Sakaridis, C.; Dai, D.; Van Gool, L. Semantic foggy scene understanding with synthetic data. *Int. J. Comput. Vis.* **2018**, *126*, 973–992. [CrossRef]
39. 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, 27–30 June 2016; pp. 3213–3223.
40. Geiger, A.; Lenz, P.; Urtasun, R. Are we ready for autonomous driving? the kitti vision benchmark suite. In Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition, Providence, RI, USA, 16–21 June 2012; IEEE: Piscataway, NJ, USA, 2012; pp. 3354–3361.
41. 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.
42. He, Z.; Zhang, L. Multi-adversarial faster-rcnn for unrestricted object detection. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 6668–6677.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.