applied sciences

MDPI

Article

CAST-YOLO: An Improved YOLO Based on a Cross-Attention Strategy Transformer for Foggy Weather Adaptive Detection

Xinyi Liu [1,2], Baofeng Zhang [1,2] and Na Liu [2,*]

[1] The School of Computer Science and Engineering, Tianjin University of Technology, No. 391 Bin Shui Xi Dao Road, Tianjin 300384, China
[2] Tianjin Key Laboratory for Control Theory and Applications in Complicated System, Tianjin University of Technology, No. 391 Bin Shui Xi Dao Road, Tianjin 300384, China
* Correspondence: liuna@email.tjut.edu.cn

Abstract: Both transformer and one-stage detectors have shown promising object detection results and have attracted increasing attention. However, the developments in effective domain adaptive techniques in transformer and one-stage detectors still have not been widely used. In this paper, we investigate this issue and propose a novel improved You Only Look Once (YOLO) model based on a cross-attention strategy transformer, called CAST-YOLO. This detector is a Teacher–Student knowledge transfer-based detector. We design a transformer encoder layer (TE-Layer) and a convolutional block attention module (CBAM) to capture global and rich contextual information. Then, the detector implements cross-domain object detection through the knowledge distillation method. Specifically, we propose a cross-attention strategy transformer to align domain-invariant features between the source and target domains. This strategy consists of three transformers with shared weights, identified as the source branch, target branch, and cross branch. The feature alignment uses knowledge distillation, to address better knowledge transfer from the source domain to the target domain. The above strategy provides better robustness for a model with noisy input. Extensive experiments show that our method outperforms the existing methods in foggy weather adaptive detection, significantly improving the detection results.

Keywords: computer vision; object detection; deep learning; domain adaptation

Citation: Liu, X.; Zhang, B.; Liu, N. CAST-YOLO: An Improved YOLO Based on a Cross-Attention Strategy Transformer for Foggy Weather Adaptive Detection. *Appl. Sci.* **2023**, *13*, 1176. https://doi.org/10.3390/app13021176

Academic Editor: Rubén Usamentiaga

Received: 6 December 2022
Revised: 10 January 2023
Accepted: 13 January 2023
Published: 15 January 2023

1. Introduction

Significant improvements have been achieved in convolutional neural network (CNN)-based object detection methods. However, the conditions necessary for these methods to achieve high accuracy are mostly limited to obeying the same feature distribution. Even with superior results on large-scale benchmark tests, there is significant performance degradation when testing in target domains with different feature distributions. The target domain may have variations in image style, lighting conditions, image quality, camera view, etc., which usually introduce a significant domain shift between the training and testing data. Although this problem can be alleviated by collecting more training data, such a solution is impractical due to the expensive and time-consuming data labeling process.

Addressing the domain shift between the source and target domain datasets is crucial for cross-domain object detection. Typically, cross-domain object detection has been used to learn robust and generalizable detectors using labeled data from the source domain and unlabeled data from the target domain. The work by [1] is a landmark study in addressing the domain shift problem in object detection. Most domain adaptive methods [2–4] have been studied based on faster regions with CNN features (Faster R-CNN) [5]. Several recent works [6,7] have proposed one-stage detector-based methods to implement cross-domain object detection while considering the computational advantages of one-stage detectors.

Therefore, to address time-sensitive practical applications, such as the foggy weather adaptive detection in road scenarios, we propose using YOLOv5 [8] as the base structure in

the framework for the study of cross-domain object detection. On the one hand, one-stage object detectors can achieve almost real-time levels and maintain comparable accuracy to two-stage object detectors. On the other hand, few studies have explored the introduction of a one-stage architecture.

In summary, to exploit the advantages of the one-stage architecture in cross-domain detection, we propose a new one-stage detection framework that designs a cross-attention strategy transformer to align features. This framework efficiently extracts the source domain's features using fully supervised learning. The samples in the source and target domains belong to the same class but come from different domains. Since the target domain samples are unlabeled, pseudo-labeling must be generated in the teacher model. We use the Mean Teacher [9] guided teacher network to detect unlabeled target images. The core concept of our work is to use a cross-attention strategy transformer, which constrains the distribution distance between the source and target domains.

Specifically, we analyze the attention weights with more focus on the domain-invariant features in different domains while ignoring the domain-specific features. Therefore, we propose a novel improved YOLO based on a cross-attention strategy transformer (CAST-YOLO) for source–target domain-invariant feature alignment.

The main contributions of this paper are as follows:

(1) A novel improved YOLO based on a cross-attention strategy transformer is proposed, named CAST-YOLO. This method uses YOLOv5 as a base architecture, which designs a transformer encoding layer and a convolutional block attention module in the detector structure to obtain rich information. This detector is based on Teacher–Student knowledge transfer.

(2) A cross-attention strategy transformer is proposed to implement the domain-invariant feature alignment between source and target. This strategy consists of three transformers with shared weights, identified as the source branch, target branch, and cross branch, respectively. Feature alignment between the cross and target branch by knowledge distillation allows better knowledge transfer from the source to the target domain.

(3) We conducted extensive experiments on public benchmarks to validate the effectiveness of our method. The experimental results demonstrate that our method achieves a significant performance improvement in the task of target domain detection.

2. Related Works

Object Detection. Early object detection methods were based on the sliding-window methods, which applied handcrafted features and classifiers on dense image grids to locate objects. However, the traditional handcrafted feature extraction method for object detection has some limitations, such as poor robustness to changing objects, high time complexity, and redundant detection windows. With the arrival of the deep convolutional neural networks, it became possible to solve the problems of traditional handcrafted feature extraction methods and improve the detection speed and accuracy of object detection. The object detection task has quickly become dominated by CNN, which can be divided into two-stage object detection [5,10,11] and one-stage object detection [12–14].

Cross-domain Object Detection. The concept of domain adaptive object detection has been raised very recently for unconstrained scenes. The purpose of cross-domain object detection is to detect objects in different domains. Research in this direction was first carried out by Chen et al. [1], who proposed a domain-adaptive Faster R-CNN that reduced the difference between image-level and instance-level distributions by embedding adversarial feature adaptation in a two-stage detection pipeline. Saito et al. [15] proposed aligning shallow local perceptual fields with deeper image-level features, i.e., strong local alignment and weak global alignment. This proposal addressed the issue of adaptability from the perspective of domain diversity. The work of [16] utilized the classification consistency of image-level and instance-level predictions with the assistance of a multilabel classification model. The work of [6] proposed a center-aware feature alignment method that enabled the discriminator to focus on features from object regions.

In contrast, several recent works have attempted to address the cross-domain object detection problem with one-stage detectors [6,7,17–19]. Ref. [6] adapted [20] to explicitly extract objectivity graphs. I3Net [7] introduced a complementary module that was specifically designed for the single-shot multibox detector (SSD) [14] architecture.

Google first proposed the Transformer model in 2017 [21]. It resulted in a large performance improvement for various tasks in natural language processing (NLP). Moreover, [22] proposed a Detection TRansformer (DETR) model for object detection that provided a real end-to-end deep learning solution. In [23], the authors proposed a new sequence feature alignment method specifically designed for the adaption of the Transformer detectors. The work of [24] designed three levels of source–target feature alignment strategies based on Transformer to improve the quality of the pseudo labels in the target domain.

In this paper, we aspire to exploit the advantages of the one-stage detector and the transformer architecture to improve the performance of the cross-domain detection model.

3. Proposed Method

3.1. Framework Overview

This section presents our proposed improved YOLO based on a cross-attention strategy transformer, named CAST-YOLO. In cross-domain object detection, the training data consist of labeled source domain images and unlabeled target domain images. The purpose is to train an object detector on the training data that can generalize to the target domain.

To exploit the advantages of the one-stage architecture in cross-domain detection, we propose a new one-stage detection framework that uses a cross-attention strategy transformer to achieve the source–target feature alignment, as shown in Figure 1. The common feature extraction structure was used in Teacher and Student models, as shown in Figure 2. In the feature extraction structure, we designed a TE-Layer (Section 3.4) based on the Transformer concept to obtain rich global and contextual information. The objective function mainly consisted of three parts, $\mathcal{L}_{sup}$ denoted the training loss of the source domain samples with their corresponding ground truth labels, and $\mathcal{L}_{unsup}$ denoted the training loss of the target domain samples with their corresponding pseudo label. $\mathcal{L}_{dis}^{adv}$ denoted the distillation loss to align the features in the source–target domain. Specifically, we propose a cross-attention strategy transformer, which implements the domain-invariant feature alignment of the source and the target with knowledge distillation (Section 3.5). In addition, we explore the effect of the channel and spatial attention modules in foggy weather adaptive detection, which further improves the discriminability of our method in the target domain by suppressing the noisy information in this domain.

We designed the feature extraction structure based on the YOLOv5 [8] architecture, which consists of a Backbone, Neck, and Detect Head, as shown in Figure 2. We added a TE-Layer and CBAM into the feature extraction structure based on the original model. The TE-Layer was used to obtain rich global and contextual information, and we added this structure into the Backbone and Neck. Extracting the attention region can help our model to resist the noise interference; therefore, we integrated the CBAM into the Neck to focus its attention on useful objects. Foggy weather adaptive detection requires a high real-time performance. We found that the average precision (AP) difference between the YOLOv5s and the YOLOv5x, YOLOv5l, and the YOLOv5m was only about 1.5%, but the computational cost was much lower than the other models. Therefore, we used YOLOv5s to pursue the best detection performance.

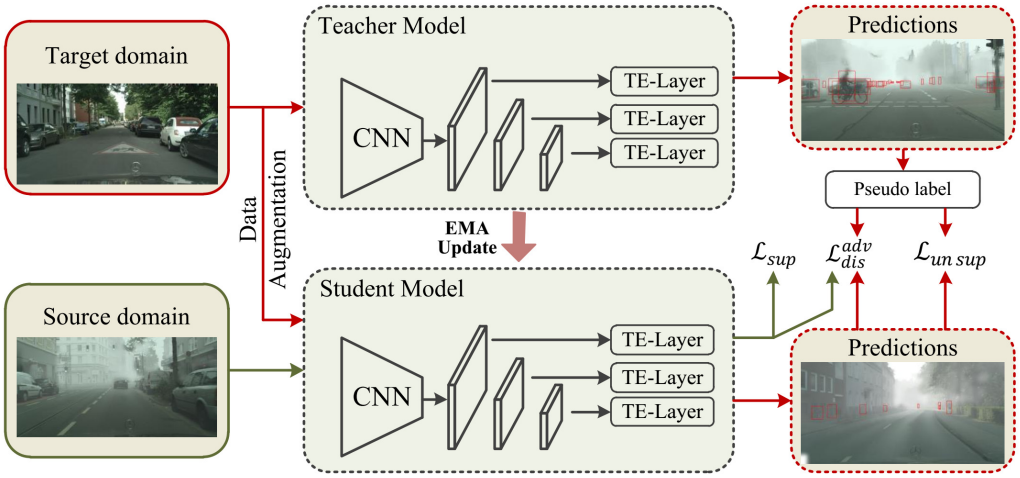


Figure 1. The overall architecture of our proposed CAST-YOLO. The framework consists of a detection model called the student model and a detection model called the teacher model. We alternate between training the student model on a supervised source domain and a pseudo-supervised target domain, and the teacher model generates pseudo labels of the target domain through exponential moving average (EMA) updates from the student model. In particular, we propose a cross-attention strategy transformer that achieves the domain-invariant feature alignment of the source–target by knowledge distillation.

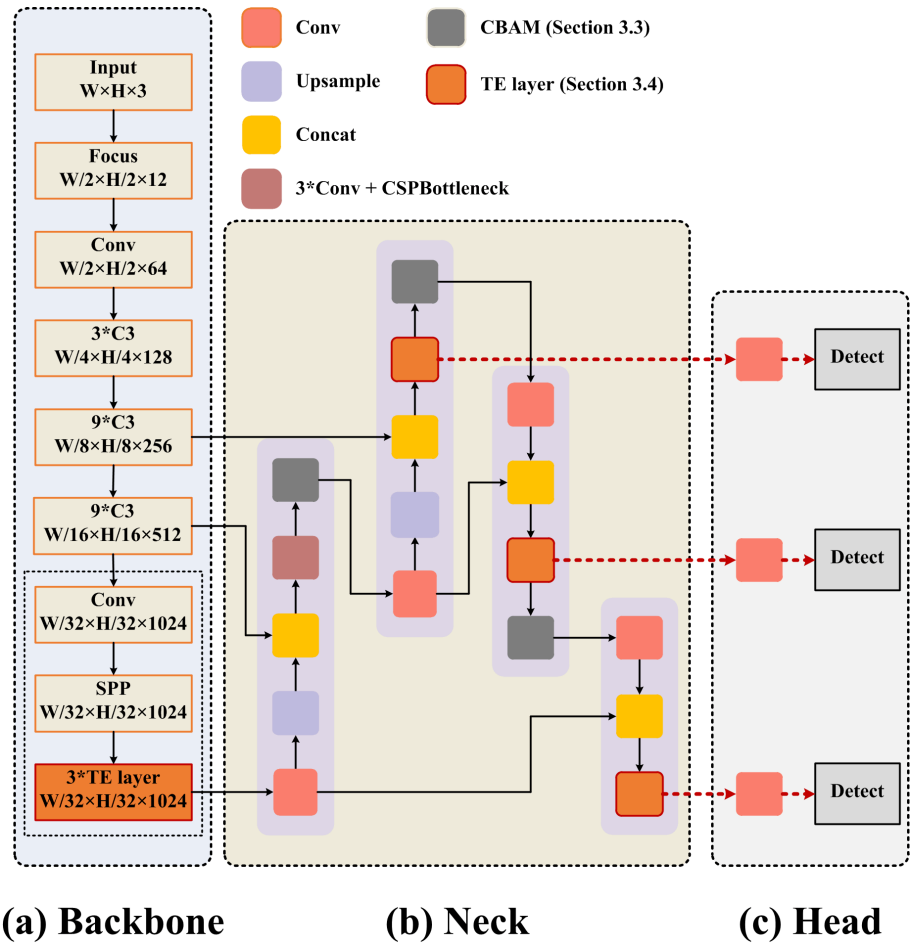


Figure 2. The architecture of feature extraction in the Teacher and Student model. (**a**) The Backbone with transformer encoder blocks at the end. (**b**) The Neck uses the PANet structure. (**c**) The Detect Heads use the feature maps from the transformer encoder blocks in the Neck.

3.2. Update Teacher from Student

In cross-domain object detection, the Teacher model is very close to the Student model, and the fact that the Teacher and Student are coupled is evident. Therefore, the Teacher model is essentially the EMA of the Student model, and their weights are tightly coupled. Our model followed the method of the Mean Teacher (MT) model [9], which consisted of two knowledge distillation structures with the same architecture (Student and Teacher).

In cross-domain object detection, the Student model is updated by backpropagation and the Teacher model is updated by the EMA weights of the Student model. Thus, the Teacher model can be considered as a collection of multiple time-wise Student models.

Specifically, we assumed that the Student and Teacher weights were defined as W_t and W'_t, respectively. In the continuous training step t, the Student weights W_t with a smoothing coefficient were used to update W'_t, as shown in Equation (1).

$$W'_t \leftarrow \alpha W'_{t-1} + (1-\alpha)W_t, \tag{1}$$

where $\alpha \in [0,1]$ denotes the smoothing coefficient. The weights of all the Teachers were updated according to Equation (1).

3.3. Channel and Spatial Attention Module

In foggy weather adaptive detection, there is noisy interference contained in extensive coverage areas, mainly from the fog itself. Extracting the attention region can help our model to resist the interference of the noisy information and focus on the useful objects. In object detection, both the category and localization accuracy of the object are important for the final detection result. The channel attention mechanism is mainly used to inform the network about what needs attention, while the spatial attention mechanism is mainly used to inform the network where attention is needed. Therefore, we propose to add the convolutional block attention module [25] into the Backbone; the structure is shown in Figure 3.

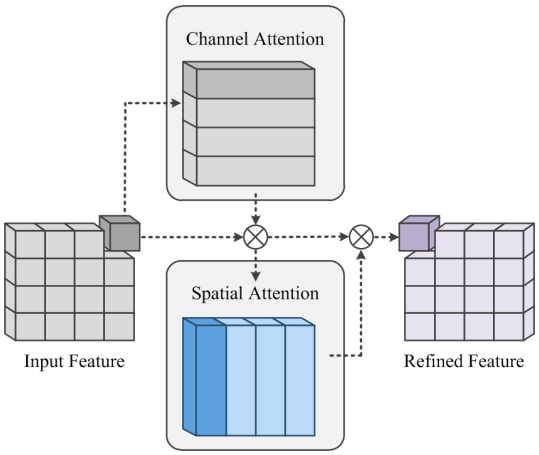


Figure 3. The channel and spatial attention module.

We defined the input features as $F \in \mathbb{R}^{C\times H\times W}$ and the final refined features as $F'' \in \mathbb{R}^{C\times H\times W}$. The channel attention was denoted as M_c, and the channel attention formula is shown in Equation (2). Firstly, the input feature was passed through two parallel *MaxPool* layers and *AvgPool* layers, which changed the feature from $C \times H \times W$ to $C \times 1 \times 1$; then, it was passed through the shared Multilayer Perceptron (MLP) module. The two postactivation results were obtained in the MLP module. The two results were summed element by element to obtain the output of the channel attention using the sigmoid function σ. The output result was multiplied by the original, which changed the feature map size back to $C \times H \times W$.

$$M_c(F) = \sigma(MLP(AvgPool(F)) + MLP(MaxPool(F))) = \sigma(W_1(W_0(F^c_{avg})) + W_1(W_0(F^c_{max}))), \tag{2}$$

where F^c_{avg} and F^c_{max} denoted the average-pooled features and max-pooled features, respectively. $W_0 \in \mathbb{R}^{C/r \times C}$ and $W_1 \in \mathbb{R}^{C \times C/r}$ were shared for both inputs, and the ReLU activation function was followed by W_0; r was the reduction ratio.

Then, the spatial attention was denoted as M_s, and its formula as shown in Equation (3). The output of the channel attention was pooled to obtain two $1 \times H \times W$ feature maps and was then spliced by a Concat operation. The spliced output was converted into a 1-channel feature map by the convolution operation f; then, the spatial attention feature was obtained using the sigmoid function σ. The output result was multiplied by the original, which changed the feature map size back to $C \times H \times W$.

$$M_s(F) = \sigma(f[AvgPool(F); MaxPool(F)] = \sigma(f[F^s_{avg}; F^s_{max}]), \tag{3}$$

where f represented a convolution operation with a filter size of 7×7.

Finally, the attention process of the convolutional block attention module can be summarized as shown in Equation (4):

$$\begin{aligned} F' &= M_c(F) \otimes F \\ F'' &= M_s(F') \otimes F', \end{aligned} \tag{4}$$

where $\otimes$ denotes the element-by-element multiplication. F denotes the intermediate feature as input; F' denotes the channel-refined feature; and F'' denotes the final refined feature.

3.4. Transformer Encoder Layer

We assumed that the transformer encoding layer would capture global information and rich contextual information. Inspired by the visual transformer [6], we replaced some convolutional and bottleneck blocks in the original YOLOv5 with the transformer encoding layer. Its structure is shown in Figure 4.

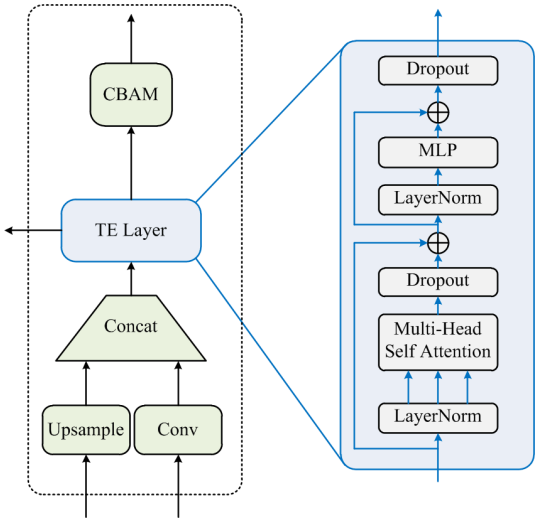


Figure 4. Transformer encoder layer.

Each transformer encoder layer contained two stages: the first stage was a multi-headed self-attention layer, and the second stage was an MLP layer. Each stage was connected with residuals. The transformer encoding layer increased the model's ability to capture different local information and had better performance on objects with high-density occlusion.

3.5. Cross-Attention Strategy Transformer

We propose a cross-attention strategy transformer to achieve domain-invariant feature alignment in the source–target domain, and the structure is shown in Figure 5. This method consisted of three transformers that shared weights. Cross-entropy was used for prediction

in the source branch (Source) and the target branch (Target), while the distillation loss was used to perform the domain-invariant feature alignment on the cross branch (Cross) and the target branch.

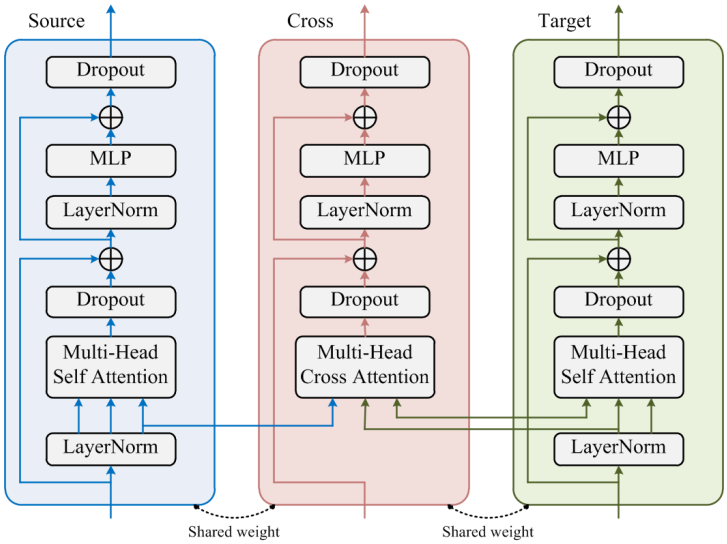


Figure 5. The transformer encoder layer.

As shown in Figure 5, the source and target domain features were fed to the source and target branches, respectively, and the multi-head self-attention module was used to learn the features. The multi-head cross-attention module was used in the cross branch, which received input from the other two branches. In the N-th layer, the query of the multi-head cross-attention module came from the query of the N-th layer of the source branch. The key and value came from the target branch. The output features of the multi-head cross-attention module were summed with the output of the $(N-1)$-th layer.

3.6. Objective Functions

In the cross-domain object detection, we had N_s source domains of labeled samples, defined as $\mathcal{D}_S = \{x_i^s, y_i^s\}_{i=1}^{N_s}$, and N_t target domains of unlabeled samples, defined as $\mathcal{D}_T = \{x_i^t\}_{i=1}^{N_t}$. x_i^s and x_i^t denoted the input samples in the source and target domains, respectively. $y_i^s = (c_i^s, b_i^s)$ denoted the labels of the corresponding input samples in the source domain; c_i^s was the category label, and b_i^s was the bounding box label.

Firstly, we trained the student model on the labeled source domain, as shown in Equation (5).

$$\mathcal{L}_{sup}(\mathcal{D}_S) = \mathcal{L}_{sup}^{cls}(x_i^s, y_{gt}) + \mathcal{L}_{sup}^{loc}(x_i^s, y_{gt}) + \mathcal{L}_{sup}^{reg} \tag{5}$$

where the source domain sample was denoted as x_i^s, and its corresponding ground truth labels were denoted as y_{gt}. $\mathcal{L}_{sup}(\mathcal{D}_S)$ denoted the object detection loss under supervision, including the classification loss, bounding box regression loss, and the regular term. $\mathcal{L}_{sup}^{cls}(x_i^s, y_{gt})$ stands for the object detection classification loss for each x_i^s source domain sample with the corresponding ground truth label y_{gt}. $\mathcal{L}_{sup}^{loc}(x_i^s, y_{gt})$ stands for the object detection bounding box regression loss for each x_i^s source domain sample with corresponding ground truth label y_{gt}. $\mathcal{L}_{sup}^{reg}$ stands for the regular term, to prevent overfitting.

As a critical element of the Teacher–Student framework, pseudo labels were generated in our method. We trained the Student model using the pseudo-labeling method in the target domain. We alternated between training the Student model on a supervised source domain and a pseudo-supervised target domain, and the Teacher model was updated by the weights of the Student model, as shown in Equation (6):

$$\mathcal{L}_{unsup}(\mathcal{D}_T) = \mathcal{L}^{cls}_{unsup}(x^t_i, \hat{y}_{gt}) + \mathcal{L}^{loc}_{unsup}(x^t_i, \hat{y}_{gt}) + \mathcal{L}^{reg}_{unsup} \quad (6)$$

where the unlabeled target domain sample was denoted as x^t_i, and its corresponding generated pseudo label was denoted as $\hat{y}_{gt}$. $\mathcal{L}_{unsup}(\mathcal{D}_T)$ is the pseudo-supervised object detection for the target domain. $\mathcal{L}^{cls}_{unsup}(x^t_i, \hat{y}_{gt})$ stands for the object detection classification loss for each x^s_i target domain sample with corresponding pseudo label $\hat{y}_{gt}$. $\mathcal{L}^{loc}_{unsup}(x^t_i, \hat{y}_{gt})$ stands for the object detection bounding box regression loss for each x^t_i target domain sample with corresponding pseudo label $\hat{y}_{gt}$. $\mathcal{L}^{reg}_{unsup}$ stands for the regular term, to prevent overfitting.

To achieve the domain-invariant feature alignment of the source and target domains, we propose a cross-attention strategy transformer to train the cross branch and target branch by the distillation loss $\mathcal{L}_{dis}$, as shown in Equation (7):

$$\mathcal{L}_{dis}(I^c, I^t) = \sum_k \mathcal{L}_{CE}(I^c_k) \log \mathcal{L}_{CE}(I^t_k), \quad (7)$$

where I^c_k and I^t_k denote the features of category k in the cross branch and the target branch, respectively. $\mathcal{L}_{CE}$ stands for the cross-entropy loss.

To summarize, the final training objective is defined as shown in Equation (8).

$$\mathcal{L} = \mathcal{L}_{sup} + \lambda_{unsup}\mathcal{L}_{unsup} + \lambda_{dis}\mathcal{L}_{dis}, \quad (8)$$

where λ_{unsup} and λ_{dis} are the tradeoff parameters.

4. Experiments and Analysis

4.1. Datasets and Implementation Details

In the scenario of weather adaptive detection, we used Cityscapes [26] as the source dataset. Cityscapes is a semantic segmentation dataset consisting of 2975 training images, 500 validation images, and 1525 testing images. Each image was annotated at the pixel level and could be used for target detection tasks after conversion. The images in the dataset were all urban scenes of different cities under normal weather. We used Foggy Cityscapes [26] as the target dataset, which was created by adding synthetic fog into the Cityscapes dataset; therefore, the annotation information was exactly the same as the original Cityscapes dataset.

Our method was implemented on Pytorch 1.8.1. All our models were trained and tested using NVIDIA RTX3090 GPUs. We used the Adam optimizer for training and 3×10^{-4} as the initial learning rate with the cosine l_r schedule. The learning rate of the last epoch decayed to 0.12 of the initial learning rate. The number of total epochs was 80. The size of the input image of our model was very large: the long side of the image was 1504 pixels, which led to the batch size being only 4.

4.2. Comparisons with the State-of-the-Art Methods

Foggy weather adaptive detection is a common and challenging task for cross-domain object detection, and the object detector must be robust in all conditions. Therefore, we evaluated the robustness of the model under foggy weather variations by converting the Cityscapes dataset to the Foggy Cityscapes dataset.

The Cityscapes dataset was used as the source domain, while the Foggy Cityscape dataset was used as the target domain. In the experiments, the model was trained with the labeled images from Cityscapes and the unlabeled images from Foggy Cityscapes. We report the testing results on the validation set of Foggy Cityscapes.

We performed Source only and Oracle experiments with YOLOv5 (marked by $*$). Source only indicates the model was trained with only the source domain data, and Oracle indicates the model was trained with labeled data from the source and target domains. ***Ours w/*** $\mathcal{L}_{cls}$ indicates the cross-entropy loss was used in the cross attention strategy to

align the source and target features. ***Ours w/trans*** indicates the model was trained with the transformer encoder layer only.

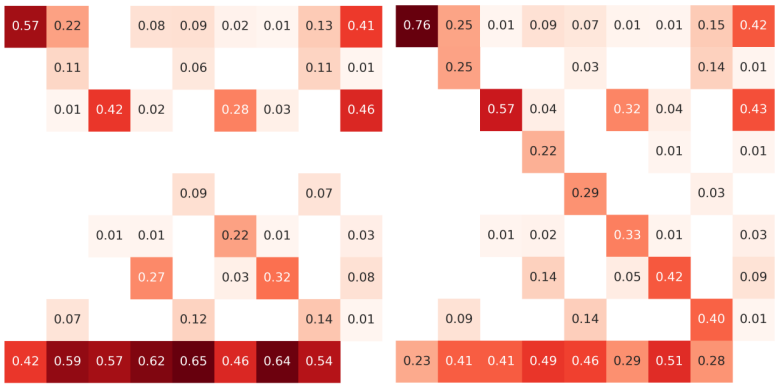
As shown in Table 1, our method outperformed other advanced methods. The state-of-the-art method SFA [23] achieved a 41.3% mAP, while our method achieved a 43.3% mAP, gaining a +2.0% improvement. This result shows that our method had a stable ability to solve foggy weather adaptive detection. Our method did not achieve the highest performance in the 'truck' and 'bus' categories. Although it did not perform better than the state-of-the-art methods in these categories, our results were very close to these methods. In addition, the AP of a few categories was relatively lower but CAST-YOLO(Ours) was substantially improved compared to the Source only detection, and the confusion matrix in Figure 6 confirms this. Meanwhile, compared with *Ours w/*$\mathcal{L}_{cls}$ and *Ours w/trans*, *CAST-YOLO(Ours)* was also substantially improved in these categories. We attributed this to the proposed cross-attention strategy transformer and the use of a more advanced convolutional block attention module.

Table 1. The results of different methods on the Foggy Cityscapes validation set for Cityscapes → Foggy Cityscapes transfer.

Methods	Detector	Person	Rider	Car	Truck	Bus	Train	Motor	Bike	mAP
DA-Faster [1]	Faster R-CNN	31.9	41.6	46.4	20.1	32.0	17.5	23.1	34.6	30.9
GPA [27]	Faster R-CNN	32.9	46.7	54.1	24.7	45.7	41.1	32.4	38.7	39.5
SFA [23]	Faster R-CNN	46.5	48.6	62.6	25.1	46.2	29.4	28.3	44.0	41.3
DIR [28]	Faster R-CNN	36.9	45.8	49.4	28.2	44.6	34.9	35.1	38.9	39.2
MS- DAYOLO [17]	YOLOv4	8.6	45.5	55.9	-	-	-	28.8	36.5	41.1
SSDA-YOLO [29]	YOLOv5-s	43.8	44.9	53.8	27.3	45.6	34.7	34.3	38.8	40.4
Source only	YOLOv5-s	37.6	38.7	59.7	14.3	33.5	11.3	2.9	31.5	28.7
Ours w/$\mathcal{L}_{cls}$	YOLOv5-s	48.0	50.0	67.5	18.8	42.8	12.5	22.2	42.1	38.0
Ours w/trans	YOLOv5-s	52.9	56.5	70.0	17.6	39.7	13.8	28.2	48.0	40.9
CAST-YOLO(Ours)	YOLOv5-s	**54.0**	**58.9**	**70.1**	21.4	43.2	19.2	28.6	51.1	**43.3**
Oracle	YOLOv5-s	51.2	49.2	71.9	40.1	57.7	56.3	40.1	42.3	51.1

4.3. Ablation Studies

In this section, we describe the investigation of the performance of the various proposed strategies. The Cityscapes converted to Foggy Cityscapes were used to evaluate our model. The Cityscapes dataset was used as the source domain, while the Foggy Cityscapes dataset was used as the target domain. For fair comparison, all experiments were performed under the same settings, and we refer the to Gain columns.



(**a**) Source only (**b**) Ours

Figure 6. Confusion matrix. The horizontal axis from left to right indicates the ground truth for 'car', 'truck', 'person', 'mbike', 'train', 'rider', 'bike', 'bus', and 'False Positive', respectively. The vertical axis from top to bottom indicates the prediction of 'car', 'truck', 'person', 'mbike', 'train', 'rider', 'bike', 'bus', and 'False Negative', respectively.

4.3.1. Loss for the Cross-Attention Strategy Transformer

We investigated the effect of different losses on the cross-attention strategy transformer. We present the results of the ablation study in Table 2. *Source only* indicates a model was trained with only source domain data. ***Ours w/*$\mathcal{L}_{cls}$** indicates the cross-entropy loss was used in the cross-attention strategy to align the source and target features, and ***Ours w/*$\mathcal{L}_{dis}$** indicates the distillation loss was used in the cross-attention strategy to align the source and target features.

Table 2. Comparison among the different loss functions on the cross-attention strategy transformer.

Methods	mAP(%)	Gain	$AP_{0.5:0.95}$(%)	Gain
Source only	28.7	+0.0%	15.7	+0.0%
Ours w/$\mathcal{L}_{cls}$	38.0	↑ +9.3%	21.8	↑ +6.1%
***Ours w/*$\mathcal{L}_{dis}$**	43.3	↑ +14.6%	25.8	↑ +10.1%

As shown in Table 2, ***Ours w/*$\mathcal{L}_{dis}$** achieved the best performance. Both its mAP and $AP_{0.5:0.95}$ outperformed those of ***Ours w/*$\mathcal{L}_{cls}$**, gaining +5.3% and +4.0% improvements, respectively. This finding illustrates that the distillation loss used in the cross-attention strategy transformer was beneficial for foggy weather adaptive detection.

4.3.2. Validity of the Proposed Module

We investigated the effect of our proposed structures. We present the results of the ablation study in Table 3. Ours w/o trans indicates the model was trained without the transformer encoder layer and convolutional block attention module. Ours w/trans indicates the model was trained with the transformer encoder layer only. Ours w/trans+attention indicates the model was trained with the cross-attention strategy transformer and the convolutional block attention module.

Table 3. The validity of the proposed module on the Cityscapes to Foggy Cityscapes scenario. '*trans*' and '*attention*' represent the transformer encoder layer and convolutional block attention module, respectively.

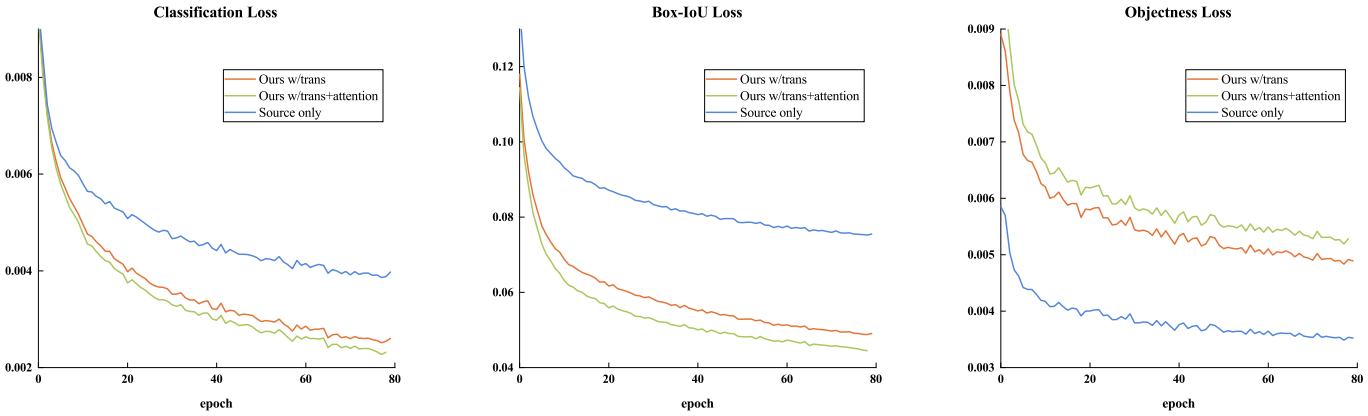
Methods	mAP(%)	Gain	$AP_{0.5:0.95}$(%)	Gain
Ours w/o trans	36.1	+0.0%	21.3	+0.0%
Ours w/trans	40.9	↑ +4.8%	23.7	↑ +2.4%
Ours w/trans+attention	43.3	↑ +7.2%	25.8	↑ +4.5%

The Ours w/trans method achieved a result of 40.9% mAP, surpassing the Ours w/o trans method and gaining a +4.8% improvement. This finding proves that the cross-attention strategy transformer was effective for foggy weather adaptive detection. Ours w/trans+attention achieved the best performance, mAP, and $AP_{0.5:0.95}$, outperforming Ours w/trans, gaining +2.4% and +2.1%, respectively. The effectiveness of the convolutional block attention module was demonstrated.

4.3.3. Scale Shift for Foggy Weather Adaptive Detection

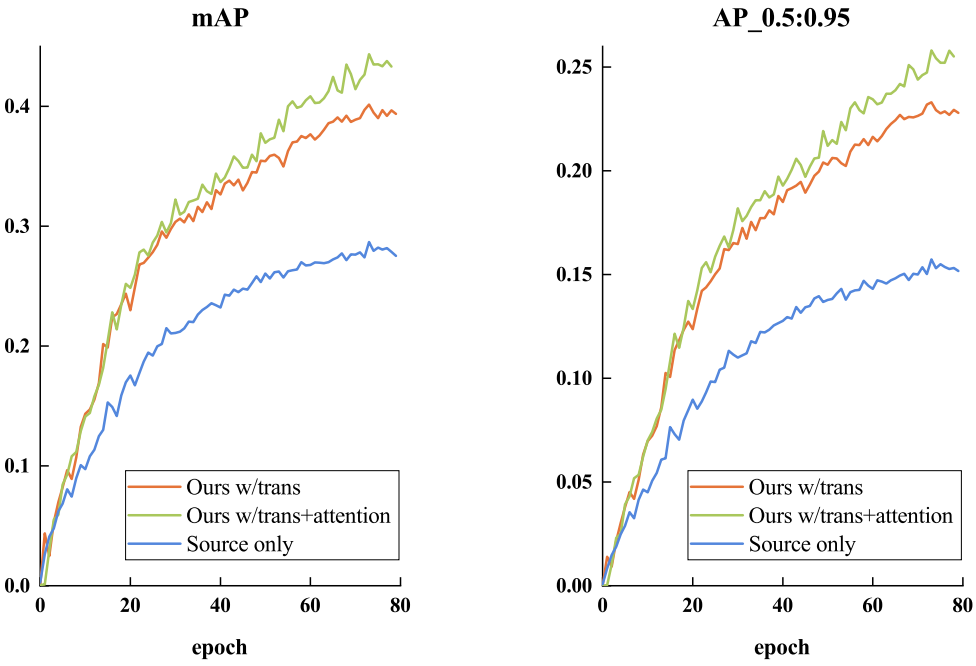
There was a potential scale shift between the source and target domain datasets. To investigate the effect of image scale on our method, we changed the size of the image in the target domain, and the scale in the source domain was fixed at 1504 pixels.

In Figures 7 and 8, the loss variation and performance comparison of different cases in the training stage are illustrated, respectively. We analyzed the effect of the scale shift on the detection performance in these three cases.



(**a**) Classification Loss (**b**) Box-IoU Loss (**c**) Objectness Loss

Figure 7. The training loss variation between the different cases. We chose three kinds of experimental setup for comparison: (1) Ours w/trans (Orange), (2) Ours w/trans+attention (Green), and (3) Source only (Blue).



(**a**) mAP (**b**) $AP_{0.5:0.95}$

Figure 8. The performance comparison between different cases in the training stage. We chose three kinds of experimental setup for comparison: (1) Ours w/trans (Orange), (2) Ours w/trans+attention (Green), and (3) Source only (Blue).

We plotted the detection performance at different image scales by changing the scales of the target domain images in Figure 9. As shown in Figure 9, by changing the scales under the same experimental conditions, the model achieved better results at all scales. Hence, our method was effective at solving the scale shift problem.

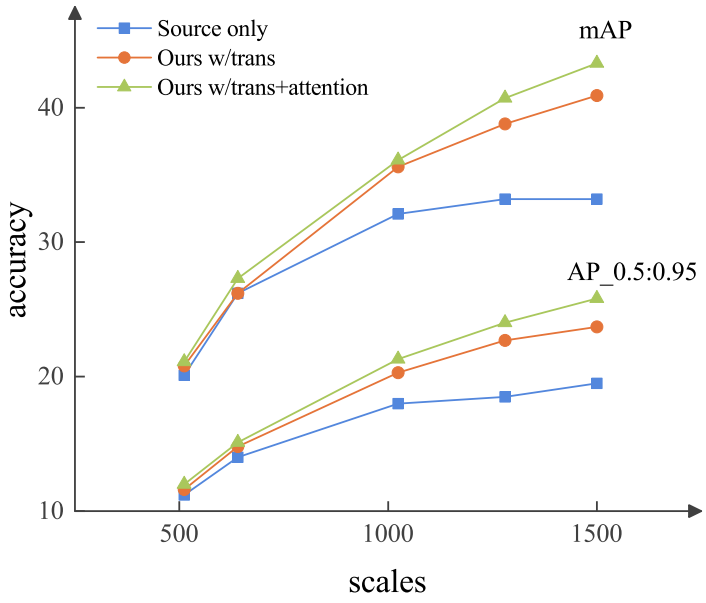


Figure 9. Scale shift for foggy weather adaptive detection. The results are from Cityscapes→Foggy Cityscapes. The scale of the image is fixed in the source domain, and we resize the different scales of image in the target domain, as shown in the X-axis.

4.4. Visualization Results

Several qualitative results are shown in Figure 10. We selected some representative images for the presentation of the test results. Figure 10 shows (a) a clear image (source domain) and (b) a foggy image (target domain), both detected with the YOLOv5 model. (c) Our proposed method (CAST-YOLO) was used for detection in foggy images.

(a) Oracle(YOLOv5) **(b)** Source(YOLOv5) **(c)** Ours(CAST-YOLO)

Figure 10. Several visualization results in the target domain.

Although YOLOv5 achieved excellent detection performance on the source domain, there were significant missing or false detections in the target domain for foggy weather adaptive detection. However, our method alleviated the limitations of YOLOv5 in foggy weather adaptive detection, which improved the classification and localization accuracy.

5. Conclusions

In this paper, we addressed the problem of foggy weather adaptive detection in cross-domain object detection by presenting a novel improved YOLO based on a cross-attention policy transformer, called CAST-YOLO. We proposed the cross-attention strategy transformer, which eliminated the effect of the domain-invariant feature shift on cross-domain object detection, improving the accuracy of foggy weather adaptive detection by a large margin. Current domain adaptive methods mostly learn feature representations from the domain or category level, which are usually too noisy for accurate feature alignment. Here, we introduced a cross-attention strategy. This strategy made the model more robust when it included noisy input and enabled better domain-invariant feature alignment. The experimental results on the Cityscapes and Foggy Cityscapes datasets demonstrated that our model achieved a performance comparable to advanced methods with improved robustness. The experimental results also showed that our detector was highly effective and advantageous in foggy weather adaptive detection.

At present, our method has only been tested in the foggy weather adaptive condition. Despite the superior detection performance achieved in foggy weather adaption, there were some limitations. For example, various tasks include cross-camera adaptation and cross-style adaptation in practical applications, but we have not yet verified the generality of our method in these adaptive conditions. Therefore, future work could consider conducting experiments under various adaptive conditions to analyze the effectiveness and generality of the proposed method.

Author Contributions: Methodology, X.L. and N.L.; investigation, N.L.; writing—original draft preparation, X.L.; writing—review and editing, B.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Research and Innovation Project for Postgraduates in Tianjin (Artificial Intelligence) grant number 2020YJSZXB08, the Youth Program of Tianjin Natural Science Foundation grant number 21JCQNJC00910, the State Key Program of Tianjin Natural Science Foundation grant number 21JCZDJC00760, and the Key Training Project for Tianjin "Project plus Team" grant number XC202054.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Acknowledgments: The authors would like to acknowledge the research support from the School of Computer Science and Engineering, the School of Electrical Engineering and Automation, and Tianjin Key Laboratory for Control Theory and Applications in Complicated System at Tianjin University of Technology.

Conflicts of Interest: The authors declare no conflict 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 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 3339–3348. [CrossRef]
2. Deng, J.; Li, W.; Chen, Y.; Duan, L. Unbiased Mean Teacher for Cross-domain Object Detection. In Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 19–25 June 2021; pp. 4091–4101. [CrossRef]

3. Mehran, K.; Arash, V.; Mani, R.; William, M. A Robust Learning Approach to Domain Adaptive Object Detection. In Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Republic of Korea, 27 October–2 November 2019; pp. 480–490. [CrossRef]
4. Yao, X.; Zhao, S.; Xu, P.; Yang, J. Multi-Source Domain Adaptation for Object Detection. In Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision (ICCV), Virtual Conference, 11–17 October 2021; pp. 3253–3262. [CrossRef]
5. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39*, 1137–1149. [CrossRef] [PubMed]
6. Cheng-Chun, H.; Yi-Hsuan, T.; Yen-Yu, L.; Ming-Hsuan, Y. Every Pixel Matters: Center-Aware Feature Alignment for Domain Adaptive Object Detector. In *Proceedings of the Computer Vision—ECCV 2020*; Springer International Publishing: Cham, Switzerland, 2020; pp. 733–748.
7. 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 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Virtual Conference, 19–25 June 2021; pp. 12571–12580. [CrossRef]
8. Glenn Jocher, K.; Nishimura, T.M.; Vilarino, R. YOLOv5. 2020. Available online: https://github.com/ultralytics/yolov5 (accessed on 12 January 2023).
9. 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, Long Beach, CA, USA, 4–9 December 2017; Volume 30.
10. Girshick, R.; Donahue, J.; Darrell, T.; Malik, J. Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation. In Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 23–28 June 2014; pp. 580–587. [CrossRef]
11. Girshick, R. Fast R-CNN. In Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV), Santiago, Chile, 7–13 December 2015; pp. 1440–1448. [CrossRef]
12. Joseph, R.; Ali, F. YOLO9000: Better, Faster, Stronger. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 6517–6525. [CrossRef]
13. Redmon, J.; Farhadi, A. YOLOv3: An Incremental Improvement. *arXiv* **2018**, arXiv:1804.02767.
14. Wei, L.; Dragomir, A.; Dumitru, E.; Christian, S.; Scott, R.; Cheng-Yang, F.; C, B.A. SSD: Single Shot MultiBox Detector. In *Proceedings of the Computer Vision—ECCV 2016*; Springer International Publishing: Cham, Switzerland, 2016; pp. 21–37.
15. Saito, K.; Ushiku, Y.; Harada, T.; Saenko, K. Strong-Weak Distribution Alignment for Adaptive Object Detection. In Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019; pp. 6949–6958. [CrossRef]
16. Xu, C.D.; Zhao, X.R.; Jin, X.; Wei, X.S. Exploring Categorical Regularization for Domain Adaptive Object Detection. In Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 14–19 June 2020; pp. 11721–11730. [CrossRef]
17. 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, AK, USA, 19–22 September 2021; pp. 3323–3327. [CrossRef]
18. Vidit, V.; Salzmann, M. Attention-based domain adaptation for single-stage detectors. *Mach. Vis. Appl.* **2022**, *33*, 65. . 10.1007/s00138-022-01320-y. [CrossRef]
19. Tian, K.; Zhang, C.; Wang, Y.; Xiang, S.; Pan, C. Knowledge Mining and Transferring for Domain Adaptive Object Detection. In Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision (ICCV), Virtual Conference, 11–17 October 2021; pp. 9113–9122. [CrossRef]
20. Kate, S.; Brian, K.; Mario, F.; Trevor, D. Adapting Visual Category Models to New Domains. In Proceedings of the Computer Vision—ECCV 2010, Heraklion, Greece, 5–11 September 2010; Springer: Berlin/Heidelberg, Germany, 2010; pp. 213–226.
21. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, L.; Polosukhin, I. Attention is All You Need. In Proceedings of the 31st International Conference on Neural Information Processing Systems, NIPS'17, Long Beach, CA, USA, 4–9 December 2017; pp. 6000–6010.
22. Carion, N.; Massa, F.; Synnaeve, G.; Usunier, N.; Kirillov, A.; Zagoruyko, S. End-to-End Object Detection with Transformers. In Proceedings of the Computer Vision—ECCV 2020: 16th European Conference, Part I, Glasgow, UK, 23–28 August 2020; pp. 213–229. ._13. [CrossRef]
23. Wen, W.; Yang, C.; Jing, Z.; Fengxiang, H.; Zheng-Jun, Z.; Yonggang, W.; Dacheng, T. Exploring Sequence Feature Alignment for Domain Adaptive Detection Transformers. *arXiv* **2021**, arXiv:2107.12636.
24. Yu, J.; Liu, J.; Wei, X.; Zhou, H.; Nakata, Y.; Gudovskiy, D.; Okuno, T.; Li, J.; Keutzer, K.; Zhang, S. MTTrans: Cross-Domain Object Detection with Mean-Teacher Transformer. *arXiv* **2022**, arXiv:2205.01643.
25. Woo, S.; Park, J.; Lee, J.Y.; Kweon, I.S. CBAM: Convolutional Block Attention Module. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018.
26. 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 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 3213–3223. [CrossRef]

27. Xu, M.; Wang, H.; Ni, B.; Tian, Q.; Zhang, W. Cross-Domain Detection via Graph-Induced Prototype Alignment. In Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 13–19 June 2020. [CrossRef]
28. Zhize, W.; Xiaofeng, W.; Tong, X.; Xuebin, Y.; Le, Z.; Lixiang, X.; Thomas, W. Domain-Invariant Proposals based on a Balanced Domain Classifier for Object Detection. *arXiv* **2022**, arXiv:2202.05941.
29. Zhou, H.; Jiang, F.; Lu, H. SSDA-YOLO: Semi-supervised Domain Adaptive YOLO for Cross-Domain Object Detection. *arXiv* **2022**, arXiv:2211.02213.