IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 11 July 2024, accepted 7 October 2024, date of publication 14 October 2024, date of current version 31 October 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3480351

RESEARCH ARTICLE

THNet: Transferability-Aware Hierarchical Network for Robust Cross-Domain Object Detection

WU SONG, SHENG REN, WENXUE TAN, AND XIPING WANG
School of Computer and Electrical Engineering, Hunan University of Arts and Science, Changde 415000, China

Corresponding author: Sheng Ren (rensheng@huas.edu.cn)

This work was supported in part by the Natural Science Foundation of Hunan Province of China under Grant 2022JJ30424, Grant 2022JJ50253, and Grant 2024JJ7317; and in part by Hunan University of Arts and Sciences Doctoral Research Initiation Project under Grant 22BSQD02.

ABSTRACT Deep learning has advanced object detection, but generalizing models from source to target domains remains a challenge due to multi-level domain drift and untransferable information. To address this, we propose a transferability-aware hierarchical domain-consistent object detector (THNet), incorporating instance-level, pixel-level, and image-level alignment subnets for robust cross-domain detection. THNet first aligns local foreground-transferable features through pixel-level adversarial learning and foreground-aware attention, then captures global domain-invariant features via image-level subnet with channel-transferable attention. Additionally, a prototype graph convolutional network alleviates instance distribution differences by maximizing inter-class distances and minimizing intra-class distances. A domain-consistent loss harmonizes training for better convergence in multi-level domain alignment. Extensive experiments demonstrate that THNet outperforms state-of-the-art methods on multiple cross-domain datasets, achieving top accuracies of 51.9%, 46.0%, 41.2%, and 51.9% across different tasks.

INDEX TERMS Cross-domain object detection, hierarchical domain alignment, domain-consistent loss, transferable attention, adversarial learning.

I. INTRODUCTION

Applying pre-built detectors to an unfamiliar domain results in a notable drop in performance because of domain shift and undefined transferable information [1]. Therefore, it is a critical and difficult task for a detector to adapt the domain shift and focus on transferable information from the source domain to an unseen target domain. To this end, in this paper, we propose a transferability-aware hierarchical domain-consistent object detector (THNet) for effective and robust cross-domain object detection.

Existing approaches for cross-domain object detection aim to minimize the distribution gap between the source and target domains, and can generally be classified into two categories. The first category of methods is to generate samples and labels [2], [3], [4]. For example, Inoue et al. [2] used pseudo labels on target domain to implement fine-tuning. Although these methods have shown remarkable performance in some scenes, ensuring the quality of generated samples and labels is challenging.

Approaches in the second category primarily emphasize feature alignment at various levels [5]. The motivation of multi-level domain alignment for cross-domain objection detection can be shown in Fig. 1. As shown in the figure, there are three levels of domain shift in cross-domain object detection, which are instance-level, pixel-level and image-level shift. The pixel-level domain shift represents the change in the distribution of image pixel values in different domains (see the histograms in Fig. 1). The image-level domain shift represents the transformation of overall image style, such as real and synthetic scenes. The instance-level domain shift represents attribute changes of object instances

The associate editor coordinating the review of this manuscript and approving it for publication was Li Zhang.

among different domains, such as the view and color of cars in Fig. 1. Most existing methods are based on two or one-level feature alignment [6], [7], [8]. For example, Chen et al. [6] proposed an adversarial loss for aligning both image-level and instance-level feature distributions. Aligning the entire process of domain shift is challenging for these two-level feature alignment methods, making it hard to achieve satisfactory cross-domain detection results. Recent works [9] have performed three-level domain alignment by using adversarial learning at each level. However, these methods usually require overwhelmingly large model and are also difficult to converge due to different learning objectives at different levels during training.

This paper introduces a transferability-aware hierarchical domain-consistent object detector (THNet) for robust cross-domain object detection to overcome these limitations. The architecture of the proposed method is shown in Fig. 2. It consists of three main components, *i.e.,* instance-level, pixel-level and image-level domain alignment subnets. This construction can help the THNet to learn domain-invariant feature space at three levels, respectively, so that obtain more robust domain-adaption performance. Furthermore, to alleviate the effect of negative transfer in three-level domain alignment, we additionally introduce two attention-based transferable modules in pixel-level and image-level feature alignment, respectively. The two modules can help to model the foreground-transferable features and alleviate untransferable features for more proper three-level domain adaption. For training, we introduce a joint multi-loss optimization objective composed of three-level losses and a domain-consistent regularization loss for optimizing the THNet properly.

The contributions of this research can be listed as follows:

1) We propose a transferability-aware hierarchical domain-consistent object detector, called THNet, to obtain multi-level transferable foreground representation and robust cross-domain object detection. Extensive experiments show that the THNet demonstrate that THNet surpasses current state-of-the-art methods, achieving top accuracies of 51.9%, 46.0%, 41.2% and 51.9% on different cross-domain detection tasks, respectively.
2) We introduce two attention-based modules, *i.e.*, the foreground-aware attention module (FAM) and channel-transferable attention module(CTM), to make the THNet have the ability to select foreground-transferable features for the pixel-level and image-level feature alignment, respectively. Both the two techniques effectively tackle the problem of negative transfer in multi-level domain adaption and improve the performance of obtaining both local and global transferable information that can also help to obtain positive transfer in the instance-level alignment.
3) In order to alleviate the hard convergence problem caused by multi-level domain adaption model with adversarial learning, a domain-consistent regularization loss is introduced to harmonize the adaptation training between pixel-level and image-level alignment.
4) To further demonstrate the robustness of our method on cross-weather object detection, a new cross-domain dataset is conducted in this paper, called Foggy Dior, which is composed of 12,225 foggy remote sensing images. We use Matlab to simulate foggy remote sensing scenarios on the Dior dataset [10]. We will release the dataset and source code after the acceptance of the paper.

II. RELATED WORK

Numerous excellent work have been proposed for domain adaptive object detection; in this section, we review those methods and discuss their shortcomings and advantages.

A. OBJECT DETECTION

With the continuous progress of convolutional neutral network (CNN) based methods in image recognition task, its application in object detection has also made amazing progress. At present, object detection methods are primarily categorized into two-stage detection and one-stage detection. As the first two-stage method, R-CNN [11] successfully introduces CNN into object detection tasks; the later Fast R-CNN [12] is implemented on the basis of R-CNN for end-to-end training. Faster R-CNN [13] introduces region proposal network to achieve faster detection. One-stage detection methods, like YOLO [14] and SSD [15], classify and localize the objects at the same time directly. However, due to the complexity and variability of real-world scenes, especially in cross-domain object detection tasks, where the source domain contains both bounding box and class annotations but the target domain lacks both, these general detection methods struggle to adapt to diverse real environments.

B. DOMAIN ADAPTATION

Domain adaptation [16], [17] has been increasingly attracting attention recently. Domain adaptation is a transfer learning method for dealing with domain transfer and lack of labels. That problem could be described as follows: The data distributions differ between the source and target domains, with the source domain having abundant labeled samples and the target domain having very limited labeled samples. For domain adaptation, a typical method is to estimate the domain gap and minimize it [18], [19]. DDC [18] is to find a kernel function, mapping both the source domain and the target domain to a Hilbert space with reproducing kernel, calculating the distance of the two domains in this space and minimizing it. Moreover, some recent researchers implement adversarial methods to obtain invariant features cross domains [19]. The improvement of DANN is CAN [20], which divides a whole CNN into several blocks, and then adds a discriminator to each block. In this way, the model ensures that the characteristics of the variables of the domain

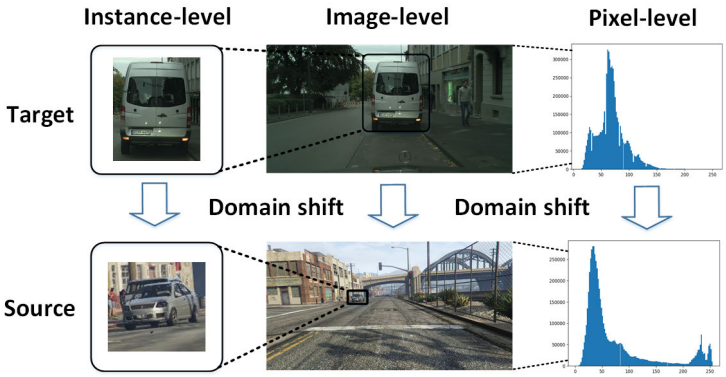


FIGURE 1. The motivation of multi-level domain alignment for cross-domain objection detection. The samples come from SIM 10k to Cityscapes.

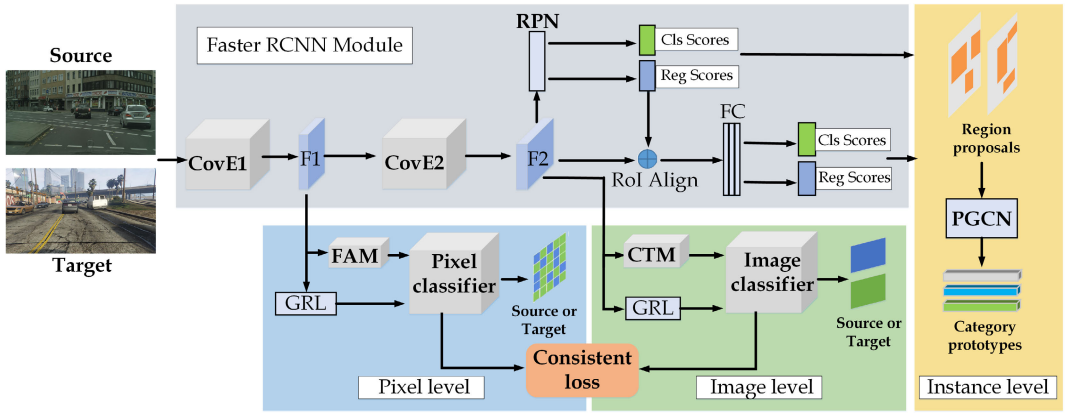


FIGURE 2. The framework of THNet. Our method performs hierarchical feature-transferable alignment (*i.e.*, instance-level, pixel-level and image-level) for mutually-reinforced domain adaption. Note: CovE1 and CovE2 represent the shallow and deep layers of the backbone, respectively, GRL represents the gradient reversal layer, FAM represents the foreground-aware attention module, CTM represents channel-transferable attention module, and PGCN represents prototype graph convolutional network.

are learned at different levels. In addition, reconstruction-based methods are improved to realize domain adaptation. DRCN [21] implements an encoder and a decoder to generate features closer to the target domain. However, this study focuses on a more challenging object detection task, where both the location and category of the object are unknown and must be accurately predicted.

C. DOMAIN ADAPTIVE OBJECT DETECTION

To mitigate performance degradation due to domain shift, existing cross-domain object detection methods are generally classified into two categories. The first category is fine-tuning methods based on generating samples and labels on target domain. Inoue et al. [2] utilized the model trained on source domain to predict pseudo labels, so that the labels can be used on target domain data for finetuning the model. RoyChowdhury et al. [3] improved the detection model's adaptability in the target domain by fine-tuning with soft labels. Wang et al. [4] and Arruda et al. [22] proposed a generator network to better understand the feature differences between the source and target domains. While these approaches have demonstrated effectiveness in certain scenarios, maintaining the quality of generated images and labels remains challenging.

The other category of cross-domain object detection methods focus on domain feature alignment in different levels, which is divided into the pixel, the image, and the instance level [5]. Most of this kind of methods are based on one or two-level alignment. For example, [8] sought to integrate category information into the domain adaptation process by introducing memory-guided attention for category-aware domain adaptation. Chen et al. [6] utilized adversarial loss to direct domain adaptation for alignment at both the image level and instance level. Reference [5] realized similar work at three levels at the same time. Xu et al. [7] constructed graph networks to extract category prototypes and induce instance-level domain alignment. Besides, Chen et al. [9] proposed adversarial learning and regularization for alignment at three levels at the same time. Despite promising results, further improvement faces several limitations. Firstly, the complexity of cross-domain object

detection makes it challenging for single or dual-level feature alignment methods to fully address domain shift throughout the detection process, thereby limiting the effectiveness of domain adaptation learning. Secondly, the training of three-level feature alignment methods is difficult to converge due to the intricate structure and numerous parameters. Moreover, most existing methods neglect to select transferable features, which may leads to negative transfer. Therefore, effectively multi-level domain adaption with transferable foreground feature selection is still an opening research problem [23].

III. PROPOSED METHOD

This study introduces a transferability-aware hierarchical domain-consistent object detector (THNet) for robust cross-domain object detection. Fig. 2 shows the overall structure of our proposed method. The THNet first employs VGG16 [24] or ResNet50 [25] as the backbone to respectively extract pixel-level and image-level features in different cross-domain tasks, then uses instance-level, pixel-level and image-level domain alignment subnets to align three-level transferable features and address domain shift in different levels. In the image-level and pixel-level domain alignment, two attention-based transferable modules, namely FAM and CTM, are proposed to alleviate negative transfer and further obtain robust transferable foreground features for more proper alignment at each level. To make the network train in an end-to-end procedure, this study propose a novel domain-consistent loss to help the THNet to optimize and thus achieve more robust cross-domain object detection performance.

A. BACKBONE FOR FEATURE EXTRACTION

To extract features of different levels for domain alignment, we first use the VGG16 [24] or the ResNet50 [25] as the backbone for feature extraction in different cross-domain tasks. In practice, the VGG16 is used for the Synthetic-to-Real task, while the ResNet50 is used for the Cross-Camera and the Normal-to-Foggy tasks. As shown in Fig. 2, following [26], we utilize the shallow layers of the backbone, named $CovE1$, to extract local features F_1 for pixel-level alignment, and the deep layers of the backbone, named as $CovE2$, to extract global features F_2 for image-level alignment. Using the shallow-layer and deep-layer features as input, we employ instance-level, pixel-level and image-level domain alignment subnets to perform domain alignment learning on them, respectively.

B. FOREGROUND-TRANSFERABLE PIXEL-LEVEL DOMAIN ALIGNMENT

For pixel-level domain alignment, we employ two adaptation modules: a foreground-aware attention module and pixel-level adversarial learning, to discover and learn pixel-level transferable features for aligning foreground-transferable information and achieving pixel-level domain adaption. The detailed architecture is shown in Fig. 3.

1) FOREGROUND-AWARE ATTENTION FOR PIXEL TRANSFER

Since local information in pixel-level extracted from the backbone is not all transferable, such as the background or noise in raw images, forcefully aligning the untransferable information can result in negative transfer [23]. However, identifying and separating the untransferable feature from transferable feature can be difficult, since it is non-trivial to define what is transferable and what is the untransferable feature. To address the problem, we employ the FAM to help to discover foreground-transferable information without defining the untransferable and transferable features explicitly. The overall pipeline of FAM can be found in Fig. 3.

More specifically, given a feature map F_1 extracted from the $CovE1$ of the backbone as input, we first apply an average-pooling layer and a max-pooling operation for feature down-sampling, and then use a 1×1 convolution and element-wise multiplication to obtain a foreground-transferable map F_1' via learning spatial attention. The obtained foreground-transferable map F_1' can be calculated as:

$$F_1' = A(F_1) \otimes F_1, \tag{1}$$

where $A(\bullet)$ represents the FAM operation and $\otimes$ represents the element-wise multiplication. Through the above process, the F_1' can effectively separate the foreground information from the background. This indicates the negative transfer could be suppressed, so that we can achieve pixel-level domain alignment in a more effective way.

2) PIXEL-LEVEL ADVERSARIAL LEARNING FOR FEATURE ALIGNMENT

With the foreground-transferable information F_1', we further use pixel-level adversarial learning to align local domain-invariant features. The pixel-level adversarial learning consists of a pixel-level domain classifier and a gradient reversal layer (GRL) [27]. In practice, the pixel-level domain classifier C_{pix} attempts to discern whether the foreground-transferable feature F_1' is from the source domain or target domain, while the $CovE1$ aims to fool the classifier. The C_{pix} and $CovE1$ are connected by the GRL to reverse the gradient flowing through the $CovE1$. Mathematically, the pixel-level adversarial loss function L_{pix} is given by:

$$L_{pix_s} = \min_{\theta_{C_{pix}}} \max_{\theta_{CovE1}} \frac{1}{n_s HW} \sum_{i=1}^{n_s} \sum_{w=1}^{W} \sum_{h=1}^{H} C_{pix}(F'_{1si})^2_{wh}, \tag{2}$$

$$L_{pix_t} = \min_{\theta_{C_{pix}}} \max_{\theta_{CovE1}} \frac{1}{n_t HW} \sum_{i=1}^{n_t} \sum_{w=1}^{W} \sum_{h=1}^{H} (1 - C_{pix}(F'_{1ti})^2_{wh}), \tag{3}$$

$$L_{pix} = \frac{1}{2}(L_{pix_s} + L_{pix_t}), \tag{4}$$

here, s and t represent the source and target domains, respectively. n is input image amount. F'_{1ti} and F'_{1si} represent the i^{th} foreground-transferable feature map of the target and source domain, respectively, with the size of $H \times W$. The coordinates of the feature maps are w and h. During

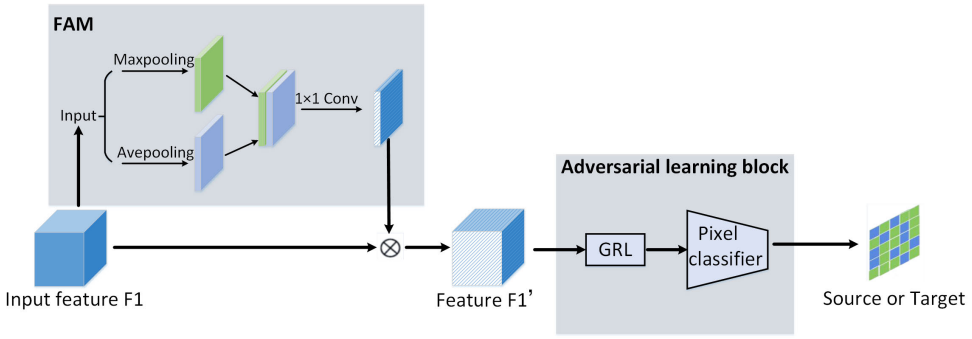


FIGURE 3. The diagram of pixel-level domain alignment subnet.

training, the network optimizes the parameters θ_{CovE1} of the backbone's shallow layers by maximizing the loss, while simultaneously minimizing the loss to optimize the parameters $\theta_{C_{pix}}$ of the domain classifier. By integrating adversarial learning with the foreground-aware attention mechanism, the pixel-level domain alignment subnet can acquire the local domain-invariant features with foreground-transferable information.

C. CHANNEL-TRANSFERABLE IMAGE-LEVEL DOMAIN ALIGNMENT

Since the deep layer of the backbone has a larger receptive field and more feature channels than the shallow layer, some of the redundant channels containing noise such as background are also not transferable and have a negative effect on the global feature alignment. To focus on aligning the transferable channel information, we introduce the CTM and image-level adversarial learning into the subnet for obtaining global transferable information and image-level domain alignment. The architecture of the CTM is detailed in Fig. 4.

1) CHANNEL-TRANSFERABLE ATTENTION FOR IMAGE TRANSFER

To avoid the negative transfer caused by forcefully aligning untransferable channels, we introduce the CTM into the image-level feature alignment subnet for discovering transferable channel information, thus weakening the untransferable channel information that includes lots of background noises. Through the process, it can help the subnet to be more robust in aligning global features of different domains.

As shown in Fig. 4, given the feature map F_2 extracted by the $CovE2$ as the input, we initially apply average-pooling and max-pooling operations for down-sampling the channel features F_2, and then use two 1×1 convolutional layers to learn the channel features for generating a channel attention descriptor. The attention weights of the attention descriptor represent the importance of the image global information, *i.e.*, a higher weight means that this channel is more transferable. With the attention descriptor, we use an element-wise multiplication to multiply the descriptor and the input feature F_2, for obtaining a channel-transferable attention map. The channel-transferable attention map F_2' is given by:

$$F_2' = T(F_2) \otimes F_2, \tag{5}$$

where $T(\bullet)$ represents the CTM operation. Obviously, the weighting operation can make the channel-transferable attention map F_2' contain more significant global transferable channel information by suppressing the redundant channel information.

2) IMAGE-LEVEL ADVERSARIAL LEARNING FOR GLOBAL FEATURE ALIGNMENT

With the discovered global transferable channel information, we further use an image-level adversarial learning with a GRL and the image-level domain classifier C_{img} to mitigate the substantial disparity among image-level global features across different domains, so that obtain global feature alignment. The image-level domain classifier is designed to determine which domain, while the GRL makes the $ConE2$ fool the classifier via the GRL. In practice, the GRL reverses the gradient during back propagation. As a result, the domain classifier can not distinguish whether the feature originates from the source or target domain, thus obtaining image-level domain-invariant features. However, in image level feature space, the distributions of features of the two domains are closer to each other, which could easily cause hard-to-classify problem. To deal with the problem, we introduce the focal loss [28], which is to focus more on the hard-to-classify samples in the target domain that are similar to the feature samples in the source domain during the training process. Here is the specific definition of the loss function,

$$L_{img_s} = -\min_{\theta_{C_{img}}} \max_{\theta_E} \frac{1}{n_s} \sum_{i=1}^{n_s} (1 - C_{img}(F_{2si}')^{\gamma}) log(C_{img}(F_{2si}')), \tag{6}$$

$$L_{img_t} = -\min_{\theta_{C_{img}}} \max_{\theta_E} \frac{1}{n_t} \sum_{i=1}^{n_t} (C_{img}(F_{2ti}')^{\gamma}) log(1 - C_{img}(F_{2ti}')), \tag{7}$$

$$L_{img} = \frac{1}{2}(L_{img_s} + L_{img_t}), \tag{8}$$

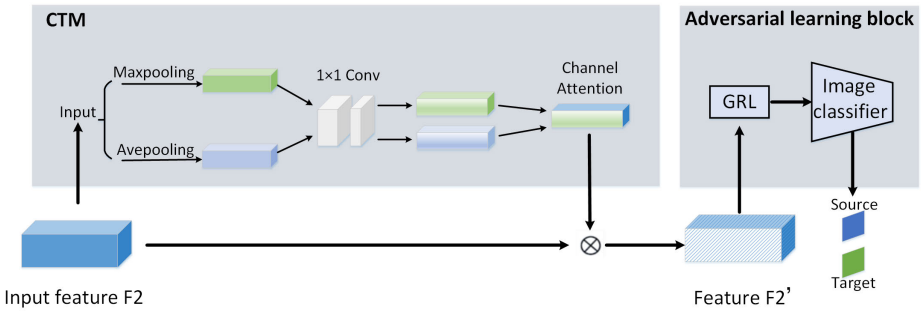


FIGURE 4. The diagram of image-level domain alignment subnet.

where γ is the weight parameter that makes the subnet focus on hard-to-classify samples during training and it is set to 5 empirically. θ_E denotes the parameters of the entire backbone. F'_{2ti} and F'_{2si} represent the i^{th} output feature maps of the $CovE2$ in the target and source domains, respectively.

D. INSTANCE-LEVEL DOMAIN ALIGNMENT

With local and global transferable features obtained by pixel-level alignment and image-level alignment subnets, region proposal network (RPN) [13] in an object detector could localize and classify the foreground proposal regions. However, due to the instance-level domain shift, such as the differences of the sizes and shapes of object instances, it is still difficult to detect accurately on the unlabeled target domain. To deal with domain shift in object instances between source and target domains, inspired from [7], we introduce a prototype graph convolutional network (PGCN) for effectively extracting and aligning category prototypes. We define that the prototype of each category is the instance-level feature vector of an object category modeled by PGCN. Fig. 5 provides the pipeline of the PGCN for instance-level domain alignment. The PGCN contains two main steps, namely, graph convolution layer for graph learning and category merging for prototype extraction.

More specifically, given the region proposals as the input, the PGCN first extracts instance-level proposal graphs via the graph convolution, and updates and outputs graphs that represent all region proposals' information of objects by training the graph convolution layer. Then, we use the category merging on all graphs of one object category to obtain the instance-level category prototypes. To fully consider the importance of different proposals, we weight the proposal regions of each category based on the corresponding confidence scores for obtaining the more robust category prototypes.

Using the extracted category prototypes, referring to [7], we employ a jointly instance-level loss L_{ins} with an intra loss and three inter losses, to train and optimize the PGCN, so that align instance-level information of category prototypes. Mathematically, the intra loss is formulated as:

$$L_{intra} = \frac{\sum_{i=1,j=1}^{0\leqslant i\neq j\leqslant n_c} \left\| c_i^s, c_j^t \right\|_2}{\sum_{i=1,j=1}^{0\leqslant i\neq j\leqslant n_c}} \tag{9}$$

where s and t represent source and target domains, respectively. c represents the category prototype of one certain category, and n_c is the number of categories. This objective function employs a L_2 regularization loss to minimize the distance between the prototypes of the same class in the source domain and the target domain. The inter loss is formulated as:

$$L_{inter(D,D')} = \frac{\sum_{i=1,j=1}^{0\leqslant i\neq j\leqslant n_c} max(0, 1 - \left\| c_i^D, c_j^{D'} \right\|_2)}{\sum_{i=1,j=1}^{0\leqslant i\neq j\leqslant n_c}} \tag{10}$$

where D and D' represent two same or different domains, respectively. This objective function is used to increase the distance between different categories across the two domains. In general, the jointly instance-level loss L_{ins} can be given by,

$$L_{ins} = \frac{1}{3}(L_{inter(s,s)} + L_{inter(t,t)} + L_{inter(s,t)}) + L_{intra}, \tag{11}$$

where $L_{inter(s,s)}$ and $L_{inter(t,t)}$ refer to the inter-class loss within the same domain, while $L_{inter(s,t)}$ represents the inter-class loss between different domains. By aligning the category prototypes with the intra and inter losses, we can make instance-level domain adaption more effectively, thus obtaining more accurate object detection in the target domain. Competitive experimental results in experimental part also demonstrate that our prototype alignment method is more effective than GRL based alignment method (about 5.4% improvement).

E. OVERALL OPTIMIZATION OBJECTIVES

For training, the THNet has three types of optimization objectives. The first one is the optimization objective of the above-mentioned three-level domain alignment losses, *i.e.*, L_{pix}, L_{img}, and L_{ins}, which are described in the corresponding sections, respectively. The second one is the proposed domain-consistent regularization loss L_{cst}. The third one is the optimization objective of the detecting loss L_{det} in Faster RCNN [13]. The total loss can be written as:

$$L_{Tol} = L_{pix} + L_{img} + L_{ins} + L_{cst} + L_{det}. \tag{12}$$

Due to the similar learning mechanisms but differing objectives in image- and pixel-level alignment, identifying and aligning domain-invariant features becomes challenging, complicating consistent feature space exploration across

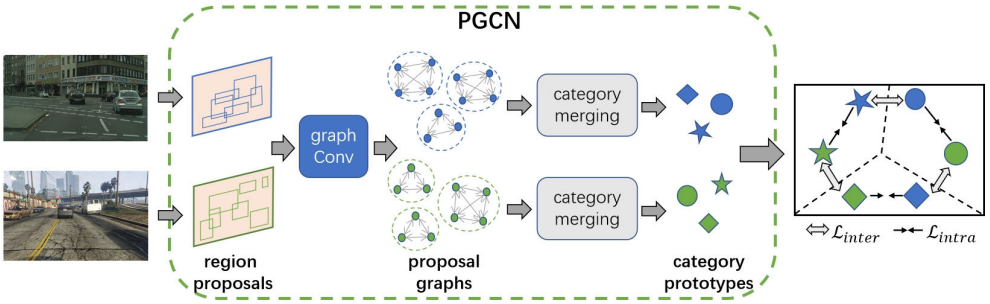


FIGURE 5. The detailed architecture of the instance-level domain alignment. The geometric shapes such as circles and triangles represent different category prototypes, respectively.

domains and hindering training convergence. To address this, we introduce a domain-consistent regularization loss L_{cst} to ensure consistent optimization in pixel- and image-level alignment:

$$L_{cst} = \beta \left\| \frac{1}{uv} \sum_{u,v} p_{pix_{uv}} - p_{img} \right\|_2, \tag{13}$$

where $p_{pix_{uv}}$ denotes the pixel-level classification probability at the pixel (u, v) of the feature map, while p_{img} represents the classification probability of the entire feature map. $\|\|_2$ is Euclidean distance used to measure the divergence between the prediction results at two levels. In this study, the empirical parameter β is set to 5. Minimizing L_{cst} aligns the prediction results of image-level and pixel-level domains to be consistent, so that make the network have a consistent optimization direction during training.

The detection loss can be written as:

$$L_{det} = \frac{1}{N_{cls}} \sum_i L_{cls}(p_i, p_i^*) + \lambda \frac{1}{N_{reg}} \sum_i p_i^* L_{reg}(t_i, t_i^*) \tag{14}$$

where L_{cls} represents the classification loss and L_{reg} denotes the regression loss. p^* is the ground-truth label and p represents the predicted classification probability. t_i represents a vector of the four parameterized coordinates of the predicted bounding box, while t_i^* corresponds to the parameterized coordinates of the ground-truth box associated with a positive anchor. We normalize the classification and regression losses with N_{cls}, N_{reg} and a balanced weight λ, and refer to Faster RCNN [13] to set them.

IV. EXPERIMENTS AND PERFORMANCE

This section presents comprehensive experimental results on three distinct cross-domain detection tasks with significant domain shift, including *Synthetic to Real* (SIM 10k [29] → Cityscapes [30]), *Cross Camera Adaptation* (KITTI [31] → Cityscapes [30]) and *Normal to Foggy* (Cityscapes →Foggy Cityscapes [32] and Dior [10]→Foggy Dior).

A. EXPERIMENTAL SETUP

The experimental setup is implemented on a 64-bit Ubuntu operating system with a single Geforce GTX2080Ti GPU using the Pytorch framework. The backbone network has been pre-trained on the ImageNet dataset [33]. The detailed experimental parameters are shown in Table 1. For the synthetic-to-real task, VGG16 [24] is used as the backbone, while ResNet50 [25] is used in cross-camera task and normal-to-foggy task. For training, the SGD optimizer is utilized with an initial learning rate of 0.001. The decay rate for the learning rate is set at 0.1, with a decay step of 5. Additionally, a warm-up strategy for the learning rate is employed during the first 200 training steps. For training and testing, 256 anchors are sampled for each image, and the ratio of positive and negative anchor samples is set to 1:1. To enhance comparison, we conducted a source-only evaluation by training solely on the source datasets and testing on the target datasets, utilizing Faster RCNN [13].

TABLE 1. Experimental parameters setting.

Parameters			Settings
Optimization	Optimizer		SGD
	Learning rate	Initial value	0.001
		Decay rate	0.1
		Decay step	5
	Warm-up steps		200
	Mini-batch size		1
	Epoch		10
Method	Focal loss gamma		5
	Anchor samples		128
	Positive-negative rate		1:1
Module	Backbone	Synthetic-to-real	VGG16 [24]
		Others	ResNet50 [25]

B. SYNTHETIC TO REAL TASK

To verify our method on the synthetic to real cross-domain task, SIM 10k is utilized as the source domain, with Cityscapes as the target domain. SIM 10k, containing 10,000 images, is collected from the game Grand Theft Auto V (GTA5). Cityscapes is a traffic scene image dataset collected by unmanned vehicle, including 2975 training images and 500 validation images. The Fig. 6 shows some examples from the two datasets. Obviously, there is huge domain shift between the source domain and the target domain, which is mainly caused by different imaging generation methods. Following the setting of other compared methods, for training, we utilized the common

car category, incorporating 10,000 source samples and 2,975 target samples; for testing, the validation split of Cityscapes, containing 500 samples, was used.

Table 2 presents a comparison between our method and state-of-the-art methods on the two datasets, including Source-only [13], DA [6], SW-DA [34], MTOR [35], GPA [7], and HDCN [36]. Our method achieved 51.9% AP for car detection, which exceeded the results of other methods. The accuracy of our method is 7.3 percent higher than SW-DA [34] and 5.3 percent higher than GPA [7]. It indicates that the THNet is more robust than the state-of-the-art methods on the synthetic-to-real domain adaptation task. Additionally, compared with HDCN [36], the AP of THNet also improved by 0.3%. It reflects the role of two-stage transferable attention for exploring transferable information.

TABLE 2. Experimental results of SIM 10k → Cityscapes on the synthetic to real cross-domain task.

Methods	car AP(%)
Source-only [13]	34.6
DA [6]	41.9
SW-DA [34]	44.6
MTOR [35]	46.6
GPA [7]	47.6
HDCN [36]	51.6
THNet(ours)	**51.9**

Fig. 7 illustrates the typical detection results of various methods from SIM 10k to Cityscapes, where green boxes indicate correct detections and red boxes indicate false detections. As shown in the Fig. 7, SW-DA, GPA and HDCN have more wrong and missed detection results than THNet, such as tiny cars. Conversely, even when dealing with tiny objects, the THNet still performs well.

C. CROSS CAMERA ADAPTATION TASK

This section investigates domain adaptation across various camera settings. The KITTI dataset serves as the source domain, while the Cityscapes dataset is employed as the target domain. The KITTI dataset is the largest computer vision dataset in the automatic driving scene in the world, which contains 7,481 training images. The Cityscapes dataset consists of 2975 training images and 500 validation images. Referring to the experimental setting in compared methods, we use its validation set for target domain testing. The data samples of the cross-camera task in the KITTI and Cityscapes datasets are shown in Fig. 8. We can find that the shapes, resolutions, as well as the weather, light, and the scenes in the data from the KITTI are significantly different from the Cityscapes.

Table 3 displays the comparison between our THNet method and state-of-the-art methods on the common car category across the two datasets. The proposed THNet method achieved the highest accuracy of 46.0%. Compared to the P-DA [37] and SC-DA [38] methods, our THNet achieved an improvement of 2.1% and 2.4%, respectively. The possible reason is that our method can effectively align transferable information for three level domain adaption by introducing two attention modules and consistent learning. Moreover, the THNet also outperformed HDCN [36] by 0.2%, which demonstrates the efficiency of transferable attention mechanism in our method.

TABLE 3. Experimental results (%) of the KITTI → Cityscapes on the cross camera adaption task.

Methods	car AP
Source-only [13]	37.6
DA [6]	41.8
SW-DA [34]	43.2
SC-DA [38]	43.6
P-DA [37]	43.9
HDCN [36]	45.9
THNet(ours)	**46.0**

Fig. 9 illustrates typical detection results from KITTI to Cityscapes, with green boxes indicating correct detections and red boxes indicating false detections. The visualization results also show that our method outperforms others obviously. As shown, the SW-DA couldn't detect tiny objects, and the localization is not accurate enough. The GPA and the HDCN detects some false results. Compared with other methods, our THNet deals with tiny objects well and there are few false detection results.

D. NORMAL TO FOGGY TASK

In this part, we verified our proposed method on the Normal to Foggy cross-weather task. To thoroughly evaluate our method, we conducted extensive experiments on two normal-to-foggy datasets. The first one is the cross-weather dateset with the traffic scene, where Cityscapes and Foggy Cityscapes datasets serve as the source and target domains, respectively. The second one is the cross-weather dateset with the remote sensing scene, where Dior and Foggy Dior datasets are used as the source and target domains, respectively. For the two scenes, the proposed THNet was tested on the validation set of the Foggy Cityscapes and testing set of Foggy Dior, respectively. Additionally, Fig. 10 presents some data samples on the two cross-weather datasets.

1) CITYSCAPES TO FOGGY CITYSCAPES

Foggy Cityscapes dataset consists of 2,975 training images and 500 validation images. The data partition of Cityscapes is the same as that of Foggy Cityscapes. For this task, Cityscapes was utilized as the source domain, while Foggy Cityscapes served as the target domain, as shown in Fig. 10 (a).

Table 4 presents a comparison between our method and other state-of-the-art techniques. Our proposed THNet achieved the highest mAP of 41.2%, and was 5.3% higher than the SC-DA and 6.1% higher than the MTOR. Compared to the second best method SCL, our THNet achieved the significant improvement of 10.3% and 7.2% on the

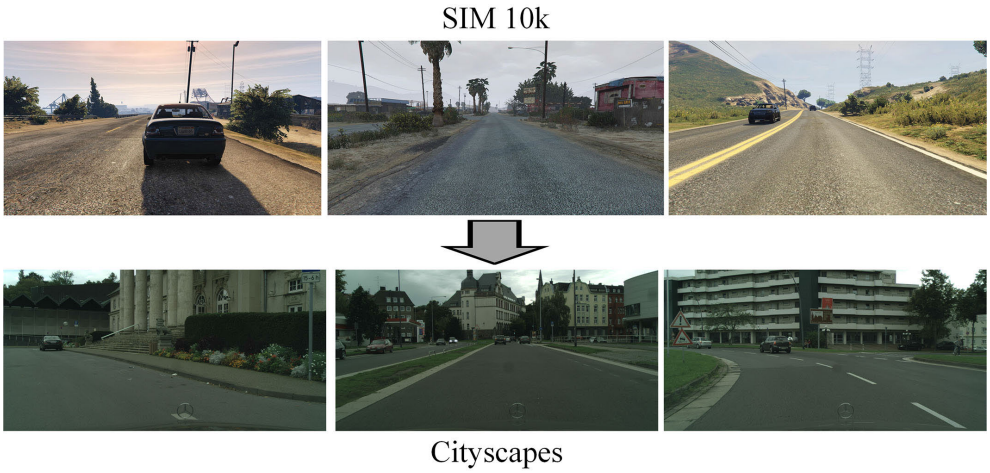
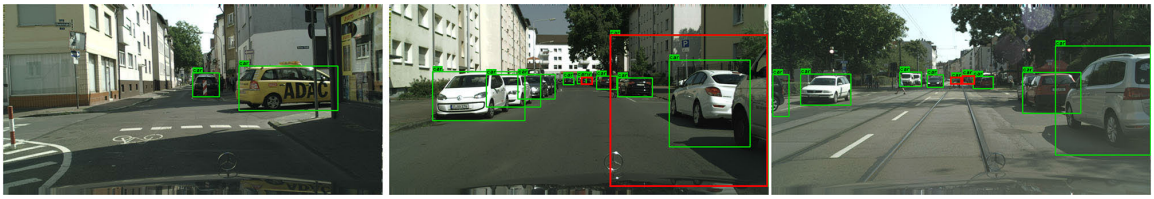
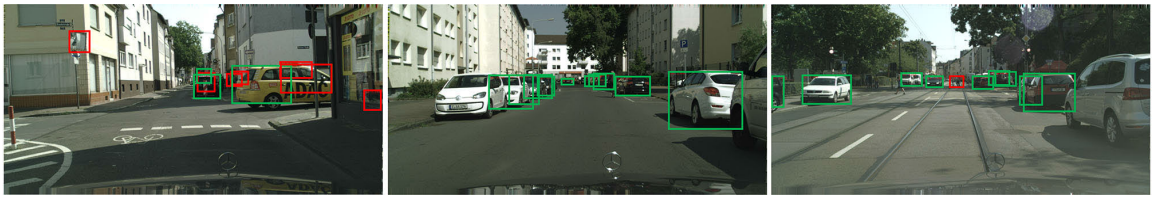


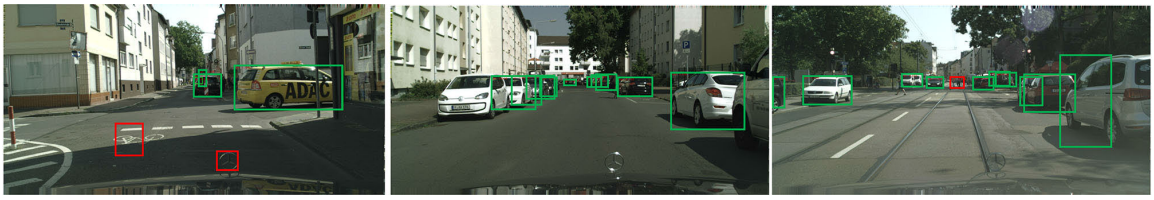
FIGURE 6. Data samples of synthetic-to-real task from the SIM 10k and Cityscapes datasets.



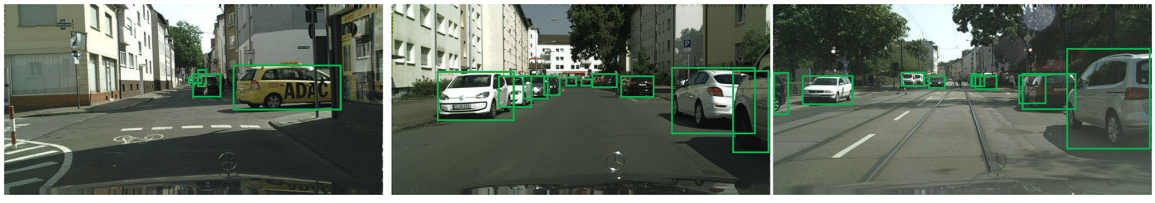
(a) SW-DA



(b) GPA



(c) HDCN



(d) THNet in this paper

FIGURE 7. Detecting results of different methods on the synthetic-to-real task. Note that green rectangular boxes indicate rightly detected targets, while red ones indicate incorrectly detected targets.

car and train categories, respectively. Moreover, the mAP of our method in all seven categories reached the best performance. The great performance shows the robustness of three-level feature alignment and the applicability of our method in the cross-weather task with the traffic scene.

Fig. 11 displays detection results of different methods on the Cityscapes to Foggy Cityscapes task, where

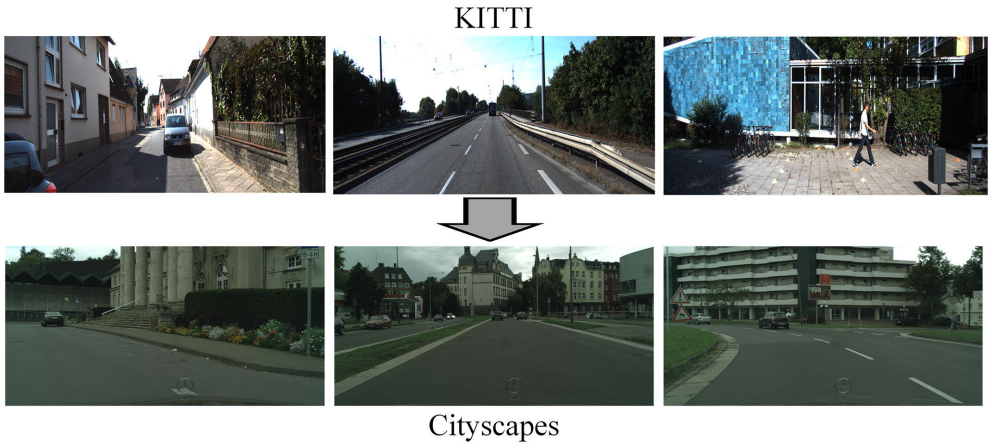
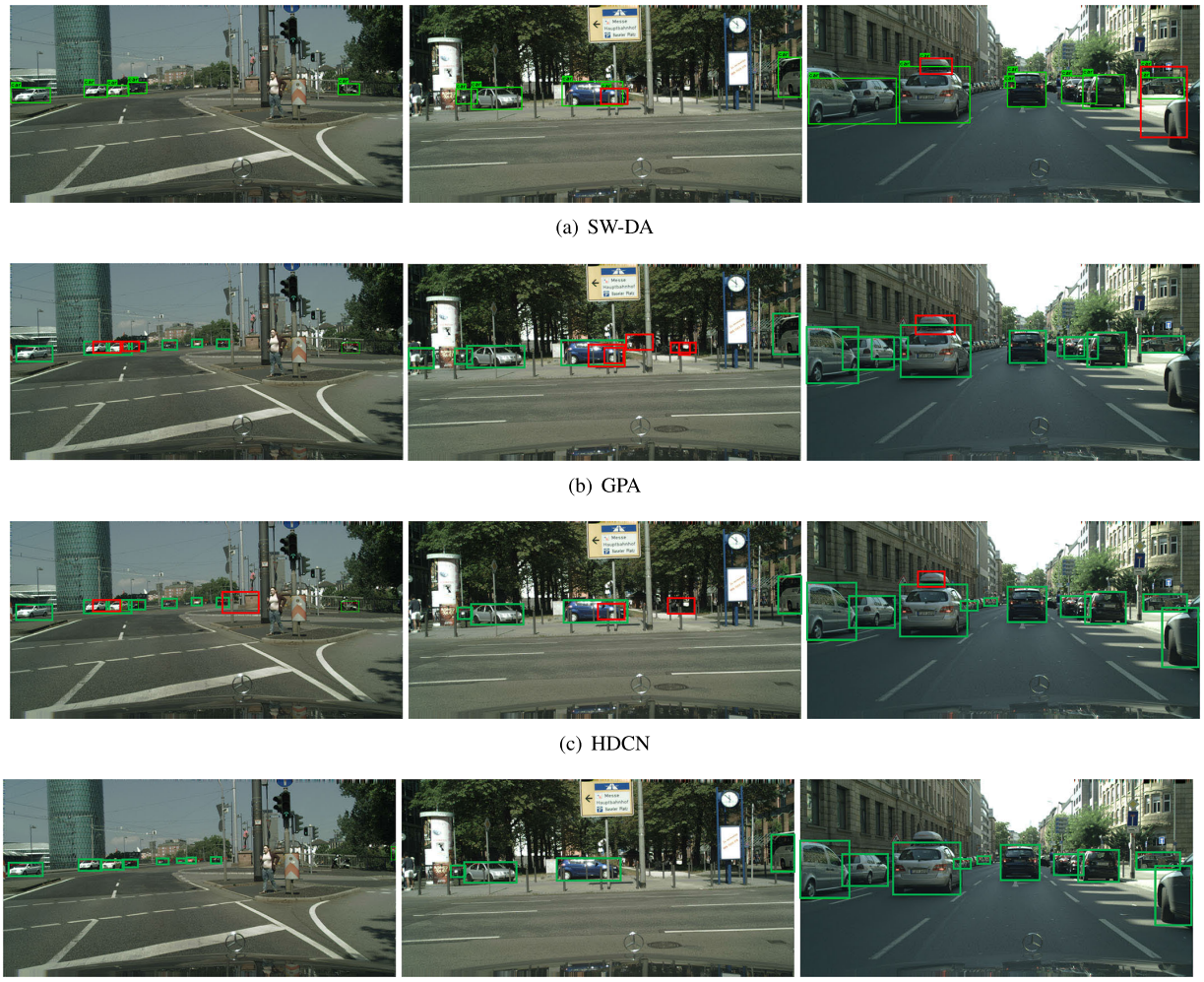


FIGURE 8. Data samples of the cross-camera task in the KITTI and Cityscapes datasets.



(a) SW-DA

(b) GPA

(c) HDCN

(d) THNet in this paper

FIGURE 9. Detecting results of different methods on the cross-camera task. Note that green rectangular boxes indicate rightly detected targets, while red ones denotes incorrectly detected targets.

green boxes indicate accurate detection results, whereas red boxes denote incorrect detection results. As shown in Fig. 11, the SW-DA missed some small objects, resulting in low accuracy. Meanwhile, the GPA and the HDCN encountered a serious false detection problem, when the objects were crowded. Compared with the other methods,

(a) Cityscapes to Foggy Cityscapes

(b) Dior to Foggy Dior

FIGURE 10. Data samples of two normal-to-foggy datasets with the traffic scene and remote sensing scene, respectively.

TABLE 4. Experimental results (%) of Cityscapes → Foggy Cityscapes on the cross-weather task.

Methods	person	rider	car	truck	bus	train	motorcycle	bicycle	mAP
Source-only [13]	26.9	38.2	35.6	18.3	32.4	9.6	25.8	28.6	26.9
DA [6]	29.2	40.4	43.4	19.7	38.3	28.5	23.7	32.7	32.0
DivMatch [39]	31.8	40.5	51.0	20.9	41.8	34.3	26.6	32.4	34.9
SW-DA [34]	31.8	44.3	48.9	21.0	43.8	28.0	28.9	35.8	35.3
SC-DA [38]	33.8	42.1	52.1	26.8	42.5	26.5	29.2	34.5	35.9
MTOR [35]	30.6	41.4	44.0	21.9	38.6	40.6	28.3	35.6	35.1
SCL [40]	31.6	44.0	44.8	30.4	41.8	40.7	**33.6**	36.2	37.9
THNet(ours)	**36.0**	**44.7**	**55.1**	**31.8**	**48.5**	**47.9**	25.6	**40.1**	**41.2**

our THNet tackled with tiny objects and crowded scenes better.

2) DIOR TO FOGGY DIOR

For further evaluation on the cross-weather object detection, we additionally performed the comparison experiments on the Dior dataset to the self-built Foggy Dior dataset. Dior is an aerospace remote sensing object detection dataset, which has 23463 images, including 11725 training images and 11738 test images with 20 categories of objects. We used Matlab to simulate fog on the 11725 training images and 500 randomly selected testing images from Dior dataset, through adding noises and changing the color channel values. We named the new dataset constructed by us as Foggy Dior, as shown in Fig. 10(b).

Table 5 presents the comparison results of our method against other state-of-the-art methods. Following the compared method [10], we use c1 to c20 to represent the 20 categories in Foggy Dior, respectively. Our method achieved the highest accuracy of 51.9% compared to the leading methods. We observed that the mAP of our THNet has a great increase of 29.2% to the GPA and 29.9% increase to that of SW-DA. Additionally, the proposed method outperformed the other methods on all the 20 categories. It further proves the effectiveness of our method on the cross-weather task even in complex remote sensing scene.

Fig. 12 displays typical detection results of different methods (*i.e.*, our method and the second best method GPA) on the Dior to Foggy Dior task, where green boxes represent correct positives. It's shown that the baseline GPA misses many tiny objects and shows bad performance on foggy remote sensing scenes. Compared with the GPA, our THNet shows great detection ability on the challenging remote sensing cross-domain task.

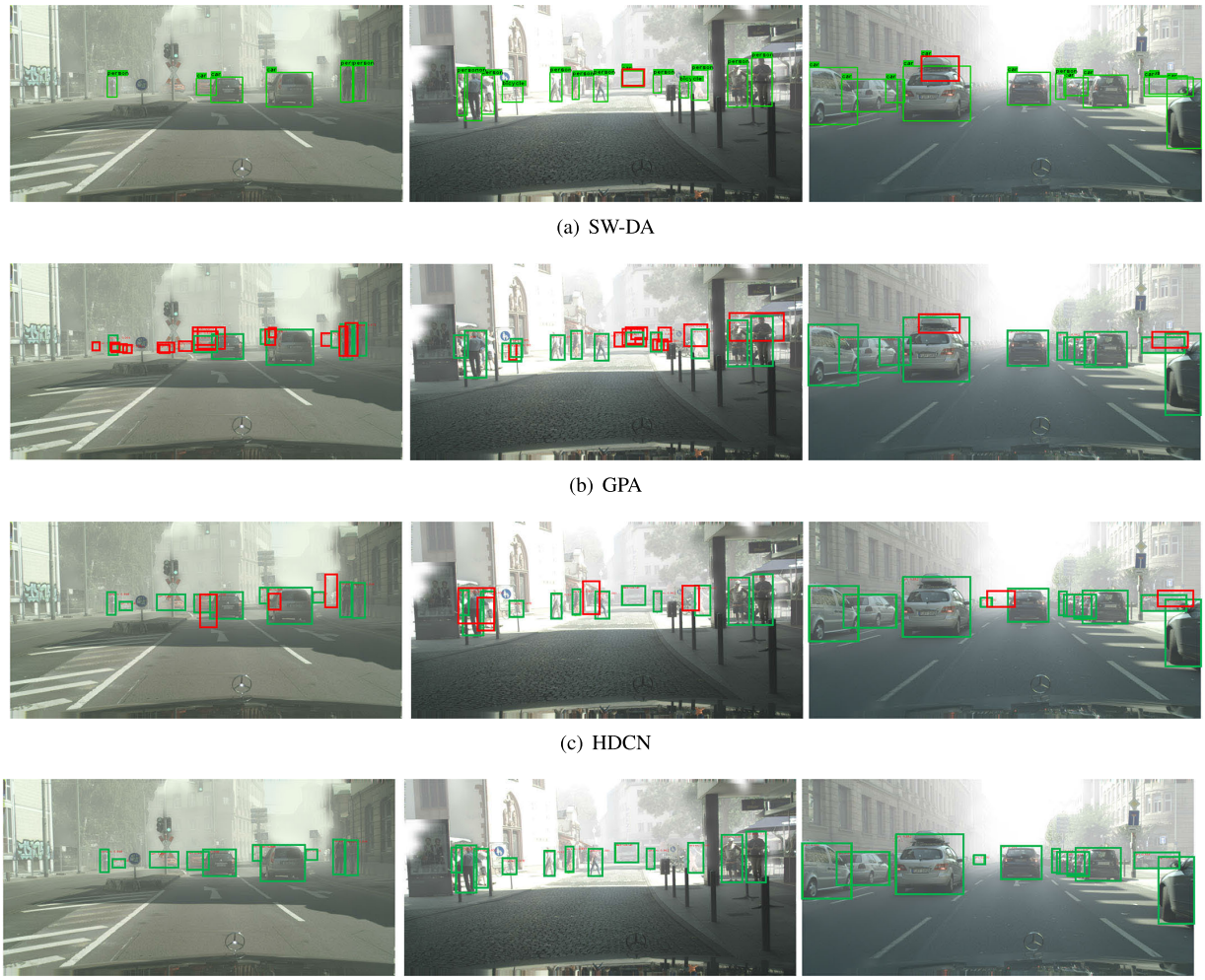
E. ABLATION ANALYSIS

1) EFFECT OF EACH MODULE

Table 6 shows the ablation experiments on the SIM 10k → Cityscapes task. We utilized the GPA detector with the ResNet50 [25] backbone as experimental baseline. GPA is a method for instance-level domain alignment. In Table 6, the detection car AP gradually increases with the sequential addition of domain alignment subnets (Ins, Img, and Pix). Adding domain-consistent regularization loss further improved AP by 1.1%. It is clear that domain-consistent loss effectively addresses the training disorder caused by different learning objectives in pixel-level and image-level alignment. Moreover, by incorporating FAM and CTM at pixel-level and image-level alignment, respectively, the detection performance in the target domain was significantly enhanced by 2.0%, demonstrating the effectiveness of the proposed transferable attention-based modules.

2) EFFECT OF DIFFERENT INSTANCE-LEVEL ALIGNMENT MECHANISM

To explore the performance of different instance-level alignment mechanism, we performed comparative experiments



(a) SW-DA

(b) GPA

(c) HDCN

(d) THNet in this paper

FIGURE 11. Comparison results of different methods on the Cityscapes → Foggy Cityscapes cross-weather task. Note that green rectangular boxes indicate rightly detected targets, while red ones presents incorrectly detected targets.

on SIM 10k → Cityscapes dataset in Table 7, including adversarial learning-based domain alignment and our PGCN. In practice, we introduced adversarial domain alignment subnet [6] instead of PGCN in the instance-level domain alignment of our THNet, and named as THNet* in Table 7. The subnet [6] of THNet* is composed of a GRL and an instance-level domain classifier to induce instance-level domain adaptation. Table shows that the car AP of our THNet with PGCN is 5.4% higher than THNet* with GRL. It indicates that the PGCN can extract and align more information of object instances.

3) EFFECT OF THE PARAMETER γ

The parameter γ of the focal loss in the image-level alignment loss shows the influence of hard-to-classify samples during training. To investigate the effect of γ, we assessed our method with varying γ on the KITTI to Cityscapes task. 13, the model attained its highest accuracy of 46.0% when γ was set to 5. We can observed that too large γ values gave too much attention to the hard-to-classify samples and over-suppressed the simple ones, which can lead to a decrease in detection accuracy. So, in the study, we set γ to 5.

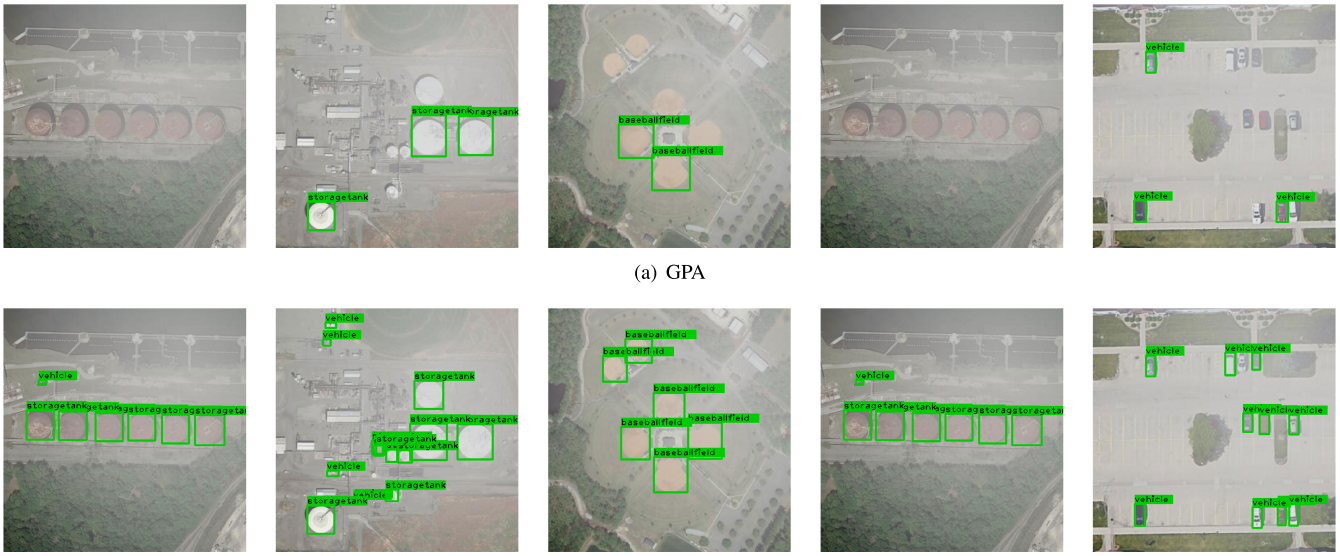
4) VISUALIZATION OF FEATURE DISTRIBUTIONS AND HEAT MAPS

Fig. 14 provides the comparison of the feature distributions with varying settings in 2D feature space using the Barnes-Hut t-SNE visualization scheme [41]. We visually compared the output features from the backbone in the source and target domains extracted by the baseline GPA and our method in the four cross-domain tasks, respectively. Due to domain-invariant foreground-transferable information learning in the three-level feature alignment, compared with the baseline GPA, our THNet method better confounded the features of two domains on the four cross-domain tasks, leading to more robust cross-domain detection results.

In addition, Fig. 15 shows attention visualization of the image-level features obtained by our method and GPA,

TABLE 5. Experimental results (%) of Dior → Foggy Dior.

Methods	c1	c2	c3	c4	c5	c6	c7	c8	c9	c10	c11	c12	c13	c14	c15	c16	c17	c18	c19	c20	mAP
Source-only [13]	13.4	4.2	38.3	23.9	0.3	32.6	22.0	27.5	16.0	28.4	16.9	10.1	8.3	0.1	20.3	0.1	29.7	0.3	0.0	4.5	14.8
DA [6]	10.3	5.2	39.7	40.8	8.1	40.9	10.6	36.4	31.3	2.1	20.3	4.7	25.5	28.6	16.3	16.7	20.9	3.7	8.1	10.0	19.0
SW-DA [34]	13.6	10.8	45.3	40.6	11.4	46.3	11.2	38.1	29.2	0.3	30.9	10.0	32.9	9.1	30.6	11.6	30.1	9.6	9.1	11.3	22.0
GPA [7]	17.5	11.6	32.1	54.3	14.3	38.8	14.5	46.2	31.5	10.5	13.0	11.0	25.3	10.7	10.6	19.2	50.1	10.5	13.6	18.9	22.7
THNet(ours)	**40.0**	**76.3**	**53.5**	**84.1**	**22.9**	**59.6**	**62.1**	**67.3**	**49.0**	**57.6**	**69.6**	**33.9**	**46.5**	**10.0**	**82.4**	**27.2**	**73.3**	**42.9**	**14.3**	**64.1**	**51.9**



(a) GPA

(b) THNet

FIGURE 12. Detecting results of Dior → Foggy Dior on the cross-weather detection task.

TABLE 6. Ablation study on the proposed THNet method. Note: Ins denotes the instance-level domain alignment subnet, and Img denotes the image-level domain alignment subnet, Pix represents the pixel-level domain alignment subnet, Con represents the domain-consistent regularization loss, and FAM and CTM represent whether using foreground-aware attention module and channel-transferable module, respectively.

Methods	Ins	Img	Pix	Con	FAM	CTM	car AP
GPA	✓						47.6
ours	✓						45.5
	✓	✓					46.4
	✓	✓	✓				48.8
	✓	✓	✓	✓			49.9
	✓	✓	✓	✓	✓		51.6
	✓	✓	✓	✓	✓	✓	51.9

TABLE 7. Comparison of different instance-level alignment mechanism.

Methods	Instance-level Adaptation	car AP(%)
THNet*	Adversarial learning [6]	46.5
THNet	PGCN	51.9

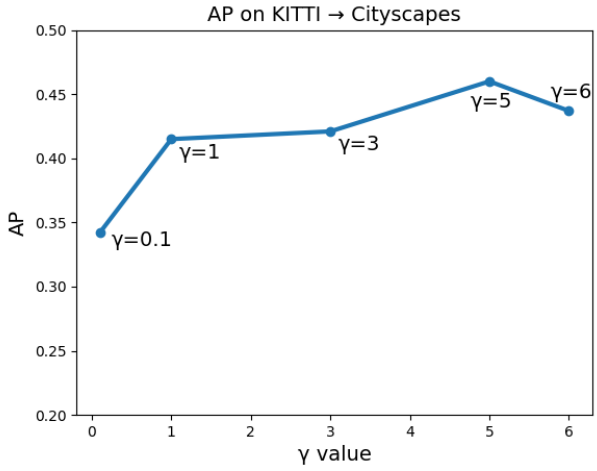


FIGURE 13. Ablation study on the parameter γ.

respectively. The experiments were implemented on the four cross-domain tasks, namely SIM 10k→Cityscapes, KITTI→Cityscapes, Cityscapes→Foggy Cityscapes, and Dior→Foggy Dior, respectively. As shown in Fig. 15, the first line is the original images, the second line is the the output feature maps of the backbone in GPA, and the third line is the image-level feature maps in our THNet, respectively. It's shown that GPA is difficult to discriminate the features of foreground and background, since the background information leaded to negative transfer during feature alignment. Obviously, we can observe that the THNet effectively focused on the foreground information (see the highlighted red in these maps). It indicates that by using the FAM and CTM can effectively alleviate the negative transfer and thus obtain robust domain adaption detection.

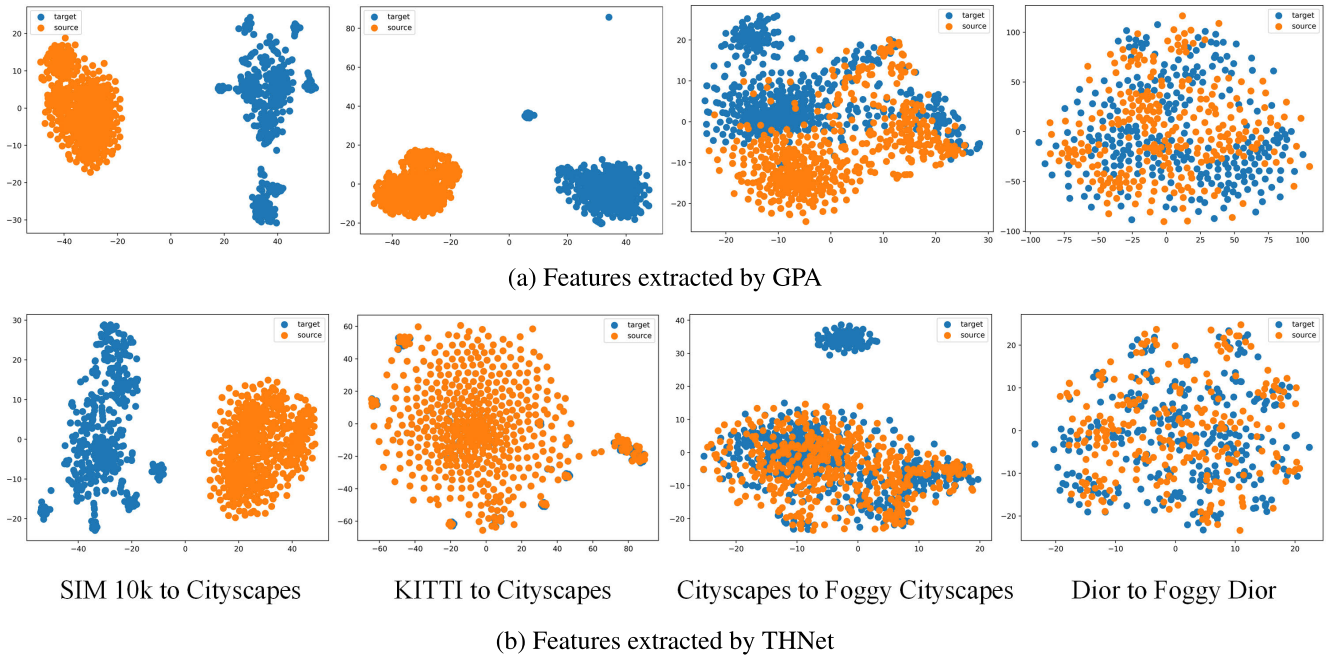
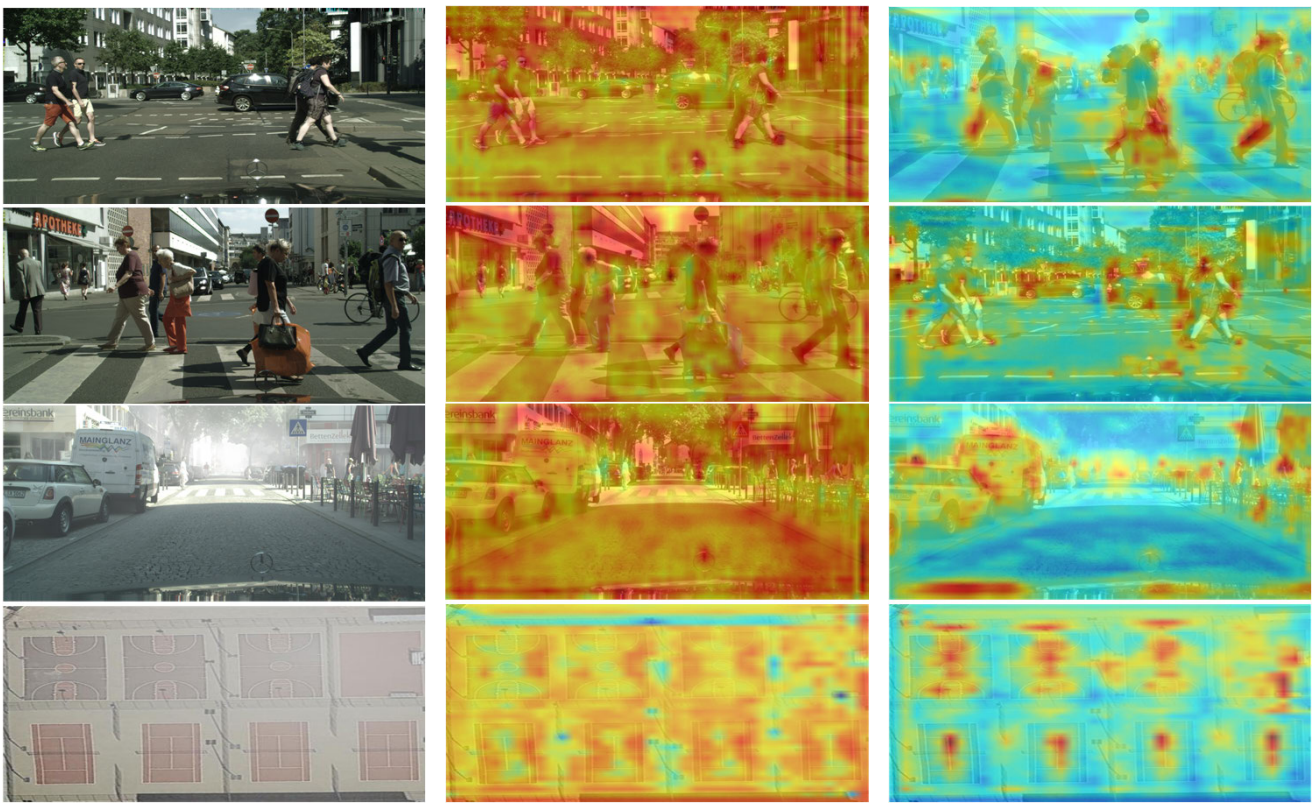


FIGURE 14. The comparison of feature distributions in source and target domains in 2D space by t-SNE feature visualization. (a) GPA, (b) our THNet.



(a) Original images (b) Attention visualization of GPA (c) Attention visualization of THNet

FIGURE 15. Attention visualization on different methods in four cross-domain tasks, including SIM 10k→Cityscapes, KITTI→Cityscapes, Cityscapes→Foggy Cityscapes, and Dior→Foggy Dior.

V. CONCLUSION AND FUTURE WORKS

This paper proposed a novel transferability-aware hierarchical domain-consistent object detection method, namely THNet, for effective and robust cross-domain object detection. The THNet consists of three main components, *i.e.*, instance-level, pixel-level and image-level domain alignment

subnets, as well as two plug-and-play attention modules, *i.e.*, foreground-aware attention and channel-transferable attention modules. Due to effectively and robustly aligning three-level features and obtaining foreground transferable representations, the proposed method achieved highly improved performance and strong robustness on several cross-domain object detection tasks. The proposed method reached the best performance on three cross-domain tasks including four challenging datasets, namely SIM 10k to Cityscapes, KITTI to Cityscapes, Cityscapes to Foggy Cityscapes, and Dior to self-built Foggy Dior, which are 51.9%, 46.0%, 41.2%, and 51.9%, respectively.

Several unresolved issues and research challenges remain in this domain. One key challenge lies in achieving a better balance between accuracy and computational efficiency in cross-domain tasks, as current methods often struggle to maintain high performance without excessive computational costs. In the future, a more efficient multi-head self-attention-based Transformer will be introduced to achieve an improved speed-accuracy trade-off.

REFERENCES

[1] T. Guo, C. P. Huynh, and M. Solh, "Domain-adaptive pedestrian detection in thermal images," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2019, pp. 1660–1664.

[2] N. Inoue, R. Furuta, T. Yamasaki, and K. Aizawa, "Cross-domain weakly-supervised object detection through progressive domain adaptation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 5001–5009.

[3] A. RoyChowdhury, P. Chakrabarty, A. Singh, S. Jin, H. Jiang, L. Cao, and E. Learned-Miller, "Automatic adaptation of object detectors to new domains using self-training," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 780–790.

[4] T. Wang, X. Zhang, L. Yuan, and J. Feng, "Few-shot adaptive faster R-CNN," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 7166–7175.

[5] Z. He and L. Zhang, "Multi-adversarial faster-RCNN for unrestricted object detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 6667–6676.

[6] Y. Chen, W. Li, C. Sakaridis, D. Dai, and L. Van Gool, "Domain adaptive faster R-CNN for object detection in the wild," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 3339–3348.

[7] M. Xu, H. Wang, B. Ni, Q. Tian, and W. Zhang, "Cross-domain detection via graph-induced prototype alignment," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 12352–12361.

[8] V. VS, V. Gupta, P. Oza, V. A. Sindagi, and V. M. Patel, "MeGA-CDA: Memory guided attention for category-aware unsupervised domain adaptive object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 4514–4524.

[9] C. Chen, Z. Zheng, X. Ding, Y. Huang, and Q. Dou, "Harmonizing transferability and discriminability for adapting object detectors," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 8866–8875.

[10] K. Li, G. Wan, G. Cheng, L. Meng, and J. Han, "Object detection in optical remote sensing images: A survey and a new benchmark," *ISPRS J. Photogramm. Remote Sens.*, vol. 159, pp. 296–307, Jan. 2020.

[11] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Mali, Jun. 2014, pp. 580–587.

[12] R. Girshick, "Fast R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 1440–1448.

[13] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 6, pp. 1137–1149, Jun. 2017.

[14] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 779–788.

[15] W. Liu, "SSD: Single shot multibox detector," in *Proc. 14th Eur. Conf.* Amsterdam, The Netherlands: Springer, Oct. 2016, pp. 21–37.

[16] M. Wang and W. Deng, "Deep visual domain adaptation: A survey," *Neurocomputing*, vol. 312, pp. 135–153, Oct. 2018.

[17] W. M. Kouw and M. Loog, "A review of domain adaptation without target labels," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 3, pp. 766–785, Mar. 2021.

[18] E. Tzeng, J. Hoffman, N. Zhang, K. Saenko, and T. Darrell, "Deep domain confusion: Maximizing for domain invariance," 2014, *arXiv:1412.3474*.

[19] Y. Ganin, E. Ustinova, H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, M. Marchand, and V. Lempitsky, "Domain-adversarial training of neural networks," *The J. Mach. Learn. Res.*, vol. 17, no. 1, pp. 2030–2096, 2016.

[20] W. Zhang, W. Ouyang, W. Li, and D. Xu, "Collaborative and adversarial network for unsupervised domain adaptation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 3801–3809.

[21] M. Ghifary, W. B. Kleijn, M. Zhang, D. Balduzzi, and W. Li, "Deep reconstruction-classification networks for unsupervised domain adaptation," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 597–613.

[22] V. F. Arruda, T. M. Paixão, R. F. Berriel, A. F. De Souza, C. Badue, N. Sebe, and T. Oliveira-Santos, "Cross-domain car detection using unsupervised image-to-image translation: From day to night," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, Jul. 2019, pp. 1–8.

[23] X. Wang, L. Li, W. Ye, M. Long, and J. Wang, "Transferable attention for domain adaptation," in *Proc. AAAI Conf. Artif. Intell.*, vol. 33, 2019, pp. 5345–5352.

[24] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*.

[25] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[26] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," in *Lecture Notes in Computer Science*. Cham, Switzerland: Springer, 2014, pp. 818–833.

[27] Y. Ganin and V. Lempitsky, "Unsupervised domain adaptation by backpropagation," in *32nd Int. Conf. Mach. Learn. (ICML)*, vol. 2, Jul. 2015, pp. 1180–1189.

[28] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2999–3007.

[29] M. Johnson-Roberson, C. Barto, R. Mehta, S. N. Sridhar, K. Rosaen, and R. Vasudevan, "Driving in the matrix: Can virtual worlds replace human-generated annotations for real world tasks?" in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, May 2017, pp. 746–753.

[30] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele, "The cityscapes dataset for semantic urban scene understanding," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 3213–3223.

[31] A. Geiger, P. Lenz, and R. Urtasun, "Are we ready for autonomous driving? The KITTI vision benchmark suite," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2012, pp. 3354–3361.

[32] C. Sakaridis, D. Dai, and L. Van Gool, "Semantic foggy scene understanding with synthetic data," *Int. J. Comput. Vis.*, vol. 126, no. 9, pp. 973–992, Sep. 2018.

[33] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 248–255.

[34] K. Saito, Y. Ushiku, T. Harada, and K. Saenko, "Strong-weak distribution alignment for adaptive object detection," 2018, *arXiv:1812.04798*.

[35] Q. Cai, Y. Pan, C.-W. Ngo, X. Tian, L. Duan, and T. Yao, "Exploring object relation in mean teacher for cross-domain detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 11449–11458.

[36] Y. Liu, Z. Liu, F. Fang, Z. Fu, and Z. Chen, "Hierarchical domain-consistent network for cross-domain object detection," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2021, pp. 474–478.

[37] H.-K. Hsu, C.-H. Yao, Y.-H. Tsai, W.-C. Hung, H.-Y. Tseng, M. Singh, and M.-H. Yang, "Progressive domain adaptation for object detection," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Mar. 2020, pp. 738–746.

[38] X. Zhu, J. Pang, C. Yang, J. Shi, and D. Lin, "Adapting object detectors via selective cross-domain alignment," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 687–696.
[39] T. Kim, M. Jeong, S. Kim, S. Choi, and C. Kim, "Diversify and match: A domain adaptive representation learning paradigm for object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 12448–12457.
[40] Z. Shen, H. Maheshwari, W. Yao, and M. Savvides, "SCL: Towards accurate domain adaptive object detection via gradient detach based stacked complementary losses," 2019, *arXiv:1911.02559*.
[41] L. van der Maaten and G. Hinton, "Visualizing data using t-SNE," *J. Mach. Learn. Res.*, vol. 9, pp. 2579–2605, Nov. 2008.

WU SONG received the Ph.D. degree in educational technology from the National Engineering Research Center for E-Learning (NERCEL), Central China Normal University, Wuhan, China, in 2019. He is an Instructor with Hunan University of Arts and Science. His research interests include educational intelligent technology, computer vision, pattern recognition, and e-learning.

SHENG REN received the Ph.D. degree in computer science and engineering from Central South University, China, in 2022. He is currently an Associate Professor with Hunan University of Arts and Sciences. His research interests include big data, image and video super-resolution, and video analysis and understanding.

WENXUE TAN received the Ph.D. degree from the College of Computer Science, Beijing University of Technology, in 2016. He is a Professor and a Senior Engineer with Hunan University of Arts and Science. He has published over 38 research papers, 19 of which was indexed by EI Compendex or SCI database, and eight of which was refereed by Chinese Science Citation Database. His current research interests include agriculture information technology, artificial intelligence, and cloud information security.

XIPING WANG received the bachelor's degree in marketing from the East China University of Technology, Jiangxi, China, in 2004. She is an Instructor with Hunan University of Arts and Science. Her current research interests include electronic commerce and information security.