Knowledge Mining and Transferring for Domain Adaptive Object Detection

Kun Tian[†‡] Chenghao Zhang[†‡] Ying Wang[†] Shiming Xiang[†‡] Chunhong Pan[†]
[†] National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences
[‡] School of Artificial Intelligence, University of Chinese Academy of Sciences
Email:{kun.tian, chenghao.zhang, ywang, smxiang, chpan}@nlpr.ia.ac.cn

Abstract

With the thriving of deep learning, CNN-based object detectors have made great progress in the past decade. However, the domain gap between training and testing data leads to a prominent performance degradation and thus hinders their application in the real world. To alleviate this problem, Knowledge Transfer Network (KTNet) is proposed as a new paradigm for domain adaption. Specifically, KTNet is constructed on a base detector with intrinsic knowledge mining and relational knowledge constraints. First, we design a foreground/background classifier shared by source domain and target domain to extract the common attribute knowledge of objects in different scenarios. Second, we model the relational knowledge graph and explicitly constrain the consistency of category correlation under source domain, target domain, as well as cross-domain conditions. As a result, the detector is guided to learn object-related and domain-independent representation. Extensive experiments and visualizations confirm that transferring object-specific knowledge can yield notable performance gains. The proposed KTNet achieves state-of-the-art results on three cross-domain detection benchmarks.

1. Introduction

Object detection is the task of identifying where and what the interested targets are in the image. It is a key component of visual perception and scene understanding, and also the basic module of many advanced visual applications, such as multi object tracking [47, 4, 53], behavior analysis [41, 2] and visual question answering [1, 39, 7]. With the development of deep learning, neural network-based models [32, 30, 31, 24, 5, 43, 42] have gradually replaced traditional machine vision methods and become the mainstream algorithm in the field of object detection. Although remarkable progress has been witnessed in modern Computer Vision systems such as autonomous driving and intelligent surveillance, the deep learning methods suffer from significant performance degradation when faced with variations of object appearance, weather and illumination.

Figure 1. Detected objects in source domain (1st row) and target domain (2nd row). The results are collected from a detector only trained on Cityscapes. Green, red, and blue boxes are true positives, false positives and false negatives. We can observe that many targets have been missed in the domain-shifted Foggy Cityscapes.

As illustrated in Figure 1, the detector trained with source domain dataset performs well on testing set with the same distribution. But if there is a severe disparity between the source and target domain, the output results are prone to missing detection (blue boxes) or false alarm (red boxes), which leads to a prominent performance degradation and hinders the deployment in real-world situations. In practice, we can mitigate this impact by establishing a task-specific dataset that covers various training samples. Unfortunately, the massive and high-quality annotation could be costly and laborious, thus being not always feasible to acquire adequate training images from new environments.

To deal with this urgent dilemma, many unsupervised domain adaptive algorithms have been proposed to improve detection accuracy in the target domain. Almost all adaptive detectors have added adversarial training modules that are actually some domain classifiers. This idea is drawn from the classification task [44]. From coarse to fine, the domain alignment modules can be divided into image-level, instance-level and pixel-level. Although adversarial training can mitigate the domain shift to a certain extent, it still has three defeats leading to deteriorated detection perfor-

mance. First, there are no restrictions on image-level feature alignment. In fact, the model should pay more attention to feature consistency of object regions. Other irrelevant information, such as background noise, does not need to be aligned. Imagine that urban streets and rural wilderness are two completely different scenes. Forcing them to have similar distribution in the feature space not only violates human intuition, but also causes the calculated loss difficult to be optimized. Second, instance-level alignment may have noise that is generated by excessive low-quality region proposals. Performing this alignment would be sensitive to inaccurate predictions. Third, adversarial training adopts minimax optimization. The gradient used to update the model parameters is alternately reversed, which will affect the stability of the training process.

Motivated by these findings, we first raise the question "Why can humans accurately recognize objects in different weather and scenes?". There are two factors worth noting. First, humans have learned the intrinsic properties of target objects. Second, humans can capture relationship knowledge between object categories, which is also independent of the domain distribution. Therefore, a new paradigm of domain adaptive detector without adversarial training is proposed. More concretely, two knowledge transfer modules are embed into the detection framework, which discover object-related knowledge from two aspects. First, we train a binary classifier shared by the source and target domain. If model can make the same classification decision for foreground and background representation in different domains, it means that the object/non-object features in source and target datasets are aligned to some extent. Second, we explore relational knowledge of object categories and explicitly constrain the consistency of relation graph between different domains, which tends to further refine the adaption process. In short, the similarity between pedestrian and rider will not decrease due to changes in weather, and the irrelevance between car and sky will not be improved due to changes in the scene. Through maintaining such object-related and domain-independent knowledge consistent, the generalization performance of detectors can be greatly enhanced.

In order to evaluate the proposed method, cross-domain testing experiments are conducted on three benchmark settings. Cityscapes [9] to Foggy Cityscapes [36] for domain adaption under different weather. Sim10k [21] to Cityscapes for synthetic domain to the real world. KITTI [14] to Cityscapes is about different scenes and cross cameras. The experimental results indicate that domain-independent knowledge mining and transferring can be used as a new paradigm for domain adaption models, which outperforms existing state-of-the-art approaches. Moreover, we also make ablation study to explore the effectiveness of each knowledge mining strategy. Qualitative visualization analysis intuitively illustrates the motivation and achievements of this paper. To sum up, our major contributions are threefold as follows.

- The proposed domain-invariant classifier can teach the model to extract common attribute knowledge of target objects, which is the basis for detector to distinguish foreground and background regions.
- We design a domain-independent category relation constraint. The generalization performance of detector is improved by explicitly constraining the consistency of category correlation between different domains.
- Comprehensive experiments and visualizations validate the effectiveness of knowledge mining and transferring. The designed method further improves the state-of-the-art level on three domain adaptation benchmarks.

2. Related work

Object detection. Object detection is a low-level computer vision task, which is regarded as a fundamental step in many advanced tasks. Most of traditional approaches [11, 10, 12] rely on manually designed features and redundant post-processes. With the thriving of deep learning, CNN-based detectors can be roughly categorized into one-stage and two-stage models. Faster R-CNN [32] designs a region proposal network to replace selective search [46], which makes the first end-to-end detection framework with faster speed and higher accuracy than its predecessors. As for one-stage detectors, YOLO [30] and SSD [27] detect objects directly from features extracted by the backbone network, without refining the classification and regression results repeatedly. In the last two years, anchor-free detectors are all the rage. Taking FCOS [43] as an example, it can predict the class and offsets of targets at each point on the feature map without preset anchors. However, all of these generic models only consider testing in the source domain, which ignores the urgent domain shift problem in the real world.

Unsupervised Domain adaptation (UDA). The goal of UDA is to generalize the model learned from labeled source domain to another unlabeled target domain, which has attracted the attention of many researchers [13, 44, 35, 22, 6, 18, 28, 26]. The classic methods can be divided into two streams. The first is based on adversarial learning, in which a classifier is proposed to improve domain-invariance on feature level [13, 44, 28] or pixel level [18, 37, 50]. Another group of methods [45, 3, 38] tries to align feature distributions through minimizing an explicitly defined domain discrepancy measurement. The specific criterion includes maximum mean discrepancy [45], $\mathcal{H}$-divergence [3] and wasserstein distance [38]. Although the study of domain adaptation has made great process, most of the above methods are usually applied to image classification tasks.

Adaptive object detectors. In the past few years, many works extend the idea of UDA to object detection task. Beginning with the Domain Adaptive Faster R-CNN [8], Chen *et al*. designed image-level and instance-level domain classifiers. Later, He *et al*. [16] and Xie *et al*. [48] added various domain classifiers for multi-layer feature adaptation, which further improve the accuracy of cross-domain testing. Satio *et al*. [34] proposed a strong-weak alignment strategy that pays more attention to similar regions while ignoring globally dissimilar images. Zhu *et al*. [55] clustered discriminative parts and aligned the corresponding local features via adversarial training. Recently, some work has put effort into achieving more accurate feature alignment. For example, Zheng *et al*. [52] presented a coarse-to-fine adaptation framework, which can progressively align deep representation. Xu *et al*. [49] designed two regularization modules to focus on object areas and hard-aligned instances. Hsu *et al*. [19] carried out pixel-level alignment and achieve better feature adaptation. He *et al*. [17] produced an asymmetric tri-way structure to enhance the transferability of detector, which consists of a chief net and an ancillary net. Since CycleGAN [54] can achieve the migration of image style, Hsu *et al*. [20] employed it for the translation from source domain images to target-like ones and then added them to the training set. For the problem of imperfect translation, Kim *et al*. [23] proposed multi-domain invariant representation learning to address domain diversification. Unlike the above methods, our KTNet does not introduce adversarial training or generative models for domain adaptation.

3. Method

In this section, an overview of the framework is first presented. Then, the intrinsic knowledge transfer module is described in details. Finally, we dig deep into the relational knowledge between various categories and constrain the consistency of class relation graph in different domains.

3.1. Overview of the framework

As illustrated in Figure 2, the training input includes a set of labeled source images $\mathcal{D}_s = \{(Train_s, B_s)\}$ and unlabeled target images $Train_t$, where B_s denotes the bounding box annotations. A shared backbone network is utilized to extract multi-level semantics feature maps and it is VGG-16 if not stated otherwise. After that, the image features are passed to three fully convolutional blocks (FCB), each of which comprises four successive convolution layers activated by ReLU function. Finally, the prediction heads utilize the local-aggregated features from FCB to produce classification scores, location offsets as well as a centerness map. Among them, the classification scores reflect category response at each position on the feature maps, regression offsets show distance to the four edges of potential bounding box for each pixel, and centerness indicates the probability that each feature point belongs to the object center. In order to correctly predict the above information, we choose Focal loss [25], IoU loss [51] and binary cross-entropy (L_{BCE}) to supervise the training of three branches. So far, the details of detection part on source domain have been introduced, which are consistent with FCOS [43]. The corresponding objective is defined as:

$$L_{det} = L_{Focal} + L_{IoU} + L_{BCE}. \tag{1}$$

To obtain more accurate bounding boxes on the testing target images $Test_t$, we propose two knowledge transfer modules to improve the generalization performance of the detector. The centerness and classification scores are used to extract object-related knowledge hidden in source and target domain. Firstly, since centerness map can highlight object centers, it is employed to attend foreground-sensitive features using Hadamard multiplication. On the other hand, we can also obtain background-related representation through reversing the activation value of centrerness map. Then, these object/non-object features are used to train a binary classifier shared by the source and target domains, as shown by blue dashed lines in Figure 2. Secondly, the classification confidence map describes the category response of each pixel position in the feature map. We first define the representation of all classes, and then calculate the category cross-correlation matrices in source, target, and cross domains as the relationship descriptors. The object-related and domain-independent knowledge is mined by explicitly constraining the consistency between category relationship graphs in different scenarios. The specific optimization objective and training processes are detailed in Section 3.2 and Section 3.3. In conclusion, the proposed knowledge transfer modules are bridges connecting the source and target domains, which can effectively capture the commonality of objects to be detected.

3.2. Domain independent classifier

From the perspective of biological cognition and objective facts, humans have a good recognition ability for target objects in different scenes, environments and weather conditions. We speculate that this may be attributed to two reasons. First, humans have learned the intrinsic knowledge of objects, which is irrelevant to the environment. For example, cars parked on city streets and trucks driving on country roads have common characteristics, such as shape and volume, which are also called shared attribute knowledge. If the model can extract object-specific features and use them to complete subsequent classification, regression and other predictions, the degradation problems that occur during cross-domain testing will be alleviated. On the contrary, if inference process is implemented by using the feature representation that is confused with scene information, then the predicts of detectors could be easily affected by the distri-

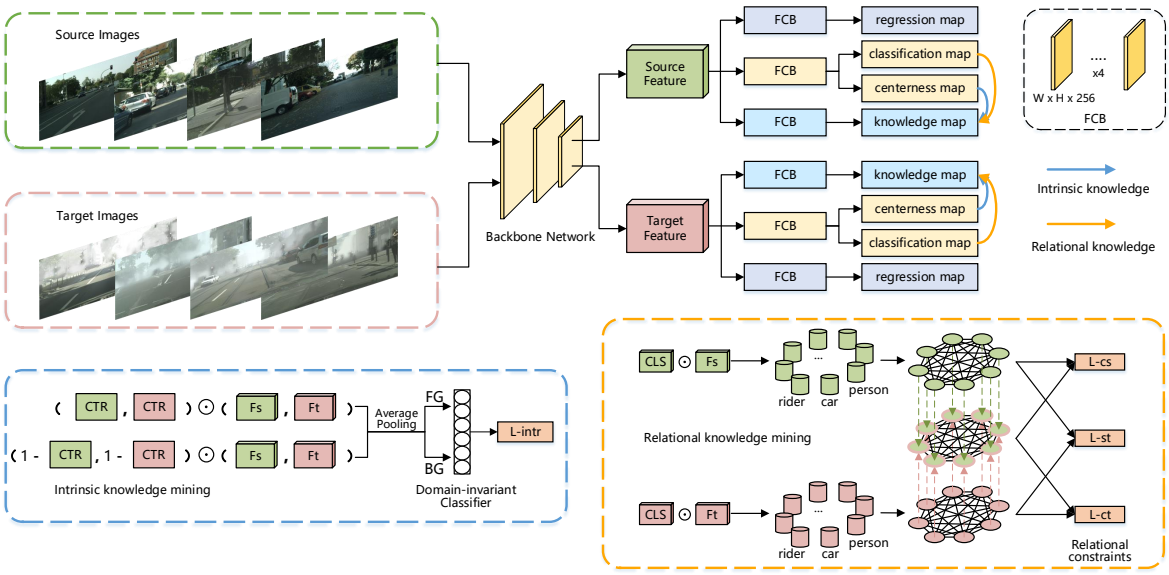

Figure 2. The schematic of our framework. Given the source and target images, a shared backbone network is used to extract image features. Settings of detection part are consistent with FCOS. F_s and F_t are knowledge maps in source and target domains. We compute Hadamard product $\odot$ of centerness maps CTR and $F_s(F_t)$ to obtain object intrinsic properties. Similar calculation is performed between classification maps CLS and $F_s(F_t)$ to extract specific category knowledge.

bution of data domain. Therefore, we build a domain independent classifier through self-supervised learning, which aims to accurately classify the foreground and background activation features from both domains. In order to generalize in source and target datasets, the model needs to extract consistent features of foregrounds and backgrounds in different scenarios, otherwise it will cause classification losses due to misjudgment. To summarize, the detector is guided to have unified predictions in both domains by training domain-invariant foreground and background classifier.

In practice, the predicted centerness map derives two kinds of mask guidance that can highlight the position coding of foreground and background pixels with great probability. Assume that the knowledge feature map extracted by FCB (the blue one) is $F \in \mathbb{R}^{H\times W\times K}$, where K represents the feature dimension and $H \times W$ corresponds to the spatial size. The mechanism of attribute knowledge mining works as follows. First, we compute Hadamard product of F and the centerness to get object-related features. As the shape of class-agnostic centerness map is $CTR \in \mathbb{R}^{H\times W}$, it needs to be duplicated for K channels to perform element-wise operation. Second, the sigmoid function is used to activate CTR, so we can get background-attentive mask by subtracting it from ones matrix. Then, the Hadamard operation is repeated with newly calculated weight map $(1 - CTR)$ to extract background features. Finally, an adaptive average pooling layer fuses the 2D representation of foreground/background into a feature vector (FG/BG), which is shown by blue dashed lines in Figure 2. The entire feature extraction process can also refer to Eq. (2):

$$\begin{aligned} FG &= \frac{\sum_{m=1}^{H}\sum_{n=1}^{W} CTR(m,n) \times F(m,n)}{H \times W}, \\ BG &= \frac{\sum_{m=1}^{H}\sum_{n=1}^{W} (1 - CTR(m,n)) \times F(m,n)}{H \times W}. \end{aligned} \tag{2}$$

In this way, each feature map can produce two semantically aggregated 1D vectors for optimizing the proposed binary classifier. During training, input images come from the source and target domain. Since source domain dataset provides supervision signals to ensure the validity of centerness map, as long as classifier has the same class prediction for background and foreground features in different domains, it can be deduced that the detector does extract some common knowledge. Moreover, the foreground/background features in source and target datasets are aligned to a certain extent, which is analyzed in the supplementary material. For simplicity, we set the classification label y_i of object and non-object features as 1 and 0, respectively. $\mathbb{P}(FG)$ is the foreground probability and the corresponding optimization objective L_{intr} is binary cross-entropy loss, which can be written as:

$$L_{intr} = -\sum_{i} y_i \log(\mathbb{P}(FG)) + (1 - y_i) \log(1 - \mathbb{P}(BG)). \tag{3}$$

Ideally, the classifier trained by Eq. (3) will predict FG in different domains as 1, and BG as 0. In other words, the detector's perception of foregrounds in both domains

should be similar. Compared with background clutter, foregrounds of different scenarios share some implicit but general attribute knowledge. Although the strategy of adversarial training is not introduced, our model is inclined to extract and maintain intrinsic characteristics of objects through knowledge mining and transferring, which reduces the domain disparity from another perspective.

3.3. Relational knowledge constraints

In addition to intrinsic attribute knowledge, humans can also capture the inherent relationship between object categories, which is not affected by external environments. Simply put, the similarity between pedestrian and rider will not decrease due to changes in weather, and the irrelevance between car and sky will not be improved due to changes in the scene. In this regard, we propose a non-parameterized constraint to teach the detector to learn domain-independent category relational knowledge as human does. A more fine-grained alignment is performed in the following three steps.

First, determine how to describe the representation of a certain category. Because the classification branch will output the prediction that can characterize the category semantics of each local region, we extract the representation from knowledge feature maps element by element with the help of classification confidence maps. As mentioned in Section 3.1, the class score map $CLS \in \mathbb{R}^{H\times W\times C}$ is obtained from the prediction head, where C is the number of categories. After being activated by sigmoid function, each pixel on CLS depicts the category response at that location. With the pixel-level classification scores, we are able to highlight the region where specific category should pay attention and activate for representation. Through utilizing the complementarity of all pixel-level features, the multi-modal information of each category can be characterized comprehensively. For instance, we multiply the confidence map CLS_{ij} of j-th category by the feature map F_i and use an average pooling layer to obtain the aggregated vector V_{ij} for i-th image. Then, the averaged V_j of all training samples is taken as the j-th category representation, which is also shown by orange dashed lines in Figure 2.

Second, define the relationship between different categories. Given input images from source and target domain, the feature descriptors of each category is regarded as one vertex of the relational graph. Then, we use two 2D feature sets P_s, P_t to maintain the class representation of source and target domain. Taking Cityscapes and Foggy Cityscapes as examples, because datasets contain eight categories, the shape of P_s and P_t is $8 \times K$, where K denotes the channel dimension of feature maps extracted from backbone network. Furthermore, the relational graph is constructed as $\mathcal{G} = \{\mathcal{V}, \mathcal{E}\}$, where $\mathcal{V}$ is the representation of each class, and $\mathcal{E}$ denotes an affinity matrix that measures the correlation between every two categories. More precisely, we calculate cosine similarity ($\mathcal{E}_{*_{ij}} = \frac{V_{*_i}\cdot V_{*_j}}{\|V_{*_i}\|_2 \cdot \|V_{*_j}\|_2}$) between two items in the feature sets to build $\mathcal{E}_*$, which is symmetric and contains object-related knowledge. As shown in Eq. (4):

$$\begin{cases} \mathcal{E}_s = \left[V_{s_0}, V_{s_1}, \cdots, V_{s_j}\right]^{\mathcal{T}} \times \left[V_{s_0}, V_{s_1}, \cdots, V_{s_j}\right] \\ \mathcal{E}_t = \left[V_{t_0}, V_{t_1}, \cdots, V_{t_j}\right]^{\mathcal{T}} \times \left[V_{t_0}, V_{t_1}, \cdots, V_{t_j}\right] \\ \mathcal{E}_c = \left[V_{s_0}, V_{s_1}, \cdots, V_{s_j}\right]^{\mathcal{T}} \times \left[V_{t_0}, V_{t_1}, \cdots, V_{t_j}\right], \end{cases} \tag{4}$$

using class representation of both domains, we can construct three relation descriptors with the shape of $j \times j$, i.e., source $\mathcal{E}_s$, target $\mathcal{E}_t$, and cross-domain $\mathcal{E}_c$ cross-correlation matrices through matrix multiplication. These relational graphs are derived to depict the correlation between categories in the source and target domains.

Finally, a new optimization objective is added to constrain the consistency of category relationship knowledge between different domains, which is calculated as follows:

$$\begin{aligned} L_{relation} &= L_{cs} + L_{ct} + L_{st} \\ &= SmoothL_1(\mathcal{E}_c - \mathcal{E}_s) \\ &+ SmoothL_1(\mathcal{E}_c - \mathcal{E}_t) + SmoothL_1(\mathcal{E}_s - \mathcal{E}_t). \end{aligned} \tag{5}$$

There are three consistency constraints L_{cs}, L_{ct}, L_{st} that penalize the inconsistent relationship with larger loss and relax the consistent ones with smaller gradient. $SmoothL_1$ function can avoid gradient explosion caused by excessive loss and gradient dispersion caused by too small value. The proposed regularization can explicitly transfer relational knowledge from source domain to the target domain. In other words, the category correlation obtained in the labeled dataset can be extended to the unlabeled data domain. Our purpose is to make the detection model focus on object-related and domain-independent knowledge, so as to improve its testing performance in the target domain. The rationale of consistency constraint is that the inherent relations between object categories should be invariant to different domain distributions.

From what has been discussed above, the overall training objective of our framework integrates the supervised detection loss L_{det} on labeled source data and two knowledge transferring losses, i.e., L_{intr} and $L_{relation}$ on both domains:

$$L_{all} = L_{det} + \lambda_1 L_{intr} + \lambda_2 L_{relation}, \tag{6}$$

where λ_1 and λ_2 are designed to balance the optimization process and the default settings are 0.5 and 1. The cooperation of L_{intr} and $L_{relation}$ leads to an adaptation that focuses on object-related knowledge, thus improving the generalization performance of original detectors. Furthermore,

we extend these knowledge transfer modules to multi layers of the backbone network, which takes feature semantics into account.

4. Experiment

In this section, we first introduce three domain shifts datasets, including normal-to-foggy, synthetic-to-real as well as cross-camera situation. Then, the detailed experimental settings are also provided. Second, we compare our KTNet with previous state-of-the-art detectors. In the end, the ablation experiments and visualization analysis intuitively demonstrate the effectiveness of our knowledge-based transfer modules.

4.1. Datasets

Cityscapes→Foggy Cityscapes. The Cityscapes dataset consists of 3,475 street scene images in different cities, which are captured by onboard cameras under normal weather condition. Foggy Cityscapes dataset is created by adding fog noise, so the images and annotations of two datasets are compatible. We regard the former and the latter as source and target domain respectively. In this experiment, we only employ the training set of Cityscapes for supervised learning and test the final model on foggy environments. There are 2,975 labeled training samples and 500 testing images.

Sim10k→Cityscapes. Sim10k dataset is obtained in the computer game scene of Grand Theft Auto V. We use 10k synthesized city scene images for training and evaluate detection accuracy on the testing set of Cityscapes. Note that only car objects are used in training, we report the results of common category among two datasets, which is the same as [8] for fair comparison.

KITTI→Cityscapes. KITTI dataset contains 14,999 images collected by vehicle-mounted cameras in real-world traffic scenes, with 7,481 for training and 7,518 for testing, which is widely used for autonomous driving research. In the experiments of cross-camera adaptation, KITTI dataset is set as the source domain and Cityscapes dataset constitutes target domain. According to the protocol of [8], we also evaluate detection accuracy on cars.

4.2. Implementation details

Considering generalization, we build the domain-adaptive model based on two kinds of backbone network, i.e., VGG-16 [40] and ResNet-101 [15], which are pre-trained on ImageNet-1K [33]. For the training process described in Section 3, two NVIDIA 1080 Ti are employed and the mini-batch per GPU is set to 4 images. More concretely, our model is trained for 24,000 iterations and stochastic gradient descent optimizer is applied to update the parameters. The learning rate is initialized to 0.005, warmed up in the first 500 iterations and decayed by a factor of 0.1 at 20,000 iterations. We use the PyTorch framework [29] to implement the training and testing process. In order to better illustrate our ideas and designs, the code is available at https://github.com/kuntian18/KTNet.

Method	person	rider	car	truck	bus	train	mbike	bike	mAP
DAF[8]	25.0	31.0	40.5	22.1	35.3	20.2	20.0	27.1	27.6
SCDA[55]	33.5	38.0	48.5	26.5	39.0	23.3	28.0	33.6	33.8
MAF[16]	28.2	39.5	43.9	23.8	39.9	33.3	29.2	33.9	34.0
SWDA[34]	29.9	42.3	43.5	24.5	36.2	32.6	30.0	35.3	34.3
MDA[48]	33.2	44.2	44.8	28.2	41.8	28.7	30.5	36.5	36.0
GACA[19]	41.9	38.7	56.7	22.6	41.5	26.8	24.6	35.5	36.0
ECR[49]	32.9	43.8	49.2	27.2	**45.1**	36.4	30.3	34.6	37.4
CTF[52]	34.0	46.9	52.1	**30.8**	43.2	29.9	**34.7**	37.4	38.6
ATF[17]	34.6	**47.0**	50.0	23.7	43.3	38.7	33.4	38.8	38.7
D&M[23]	30.8	40.5	44.3	27.2	38.4	34.5	28.4	32.2	34.6
PDA[20]	36.0	45.5	54.4	24.3	44.1	25.8	29.1	35.9	36.9
KTNet	**46.4**	43.2	**60.6**	25.8	41.2	**40.4**	30.7	**38.8**	**40.9**

Table 1. Experimental results (%) on adaptation from Cityscapes to Foggy Cityscapes.

Method	Sim10k→C	KITTI→C
DAF[8]	39.0	38.5
SWDA[34]	40.1	37.9
MAF[16]	41.1	41.0
MDA[48]	42.8	-
ATF[17]	42.8	42.1
SCDA[55]	43.0	42.5
CTF[52]	43.8	-
PDA[20]	43.9	-
GACA[19]	49.0	43.2
KTNet	**50.7**	**45.6**

Table 2. Experimental results (%) of adapting Sim10k/KITTI to Cityscapes (C). Average precision is evaluated on Car category.

4.3. Comparison with state-of-the-art methods

In this section, we conduct experiments under three cross-domain settings and make comparison to previous state-of-the-art models, including feature-level adaption: DAF [8], SCDA [55], MAF [16], SWDA [34], MDA [48], GACA [19], ECR [49], CTF [52], ATF [17] and pixel-level adaption: D&M [23], PDA [20].

Normal to Foggy. The experiments from normal images to foggy images are reported in Table 1. The KTNet obtains 40.9% mAP detection accuracy that surpasses all the counterparts on weather transfer task. Compared with the best feature-level adaption method, KTNet has higher accuracy. Especially for person and car categories, the average precision has improved by 4.5% and 3.9%. Despite not using style transfer algorithm for data augmentation, our approach

Method	backbone	person	rider	car	truck	bus	train	mbike	bike	mAP
Source-only	VGG-16	26.3	24.9	31.4	6.2	15.4	6.5	10.3	23.9	18.1
+IK	VGG-16	30.1	29.8	36.7	5.5	18.7	1.7	14.9	27.3	20.6
+RK	VGG-16	43.3	40.7	59.1	24.6	39.8	28.2	30.6	35.9	37.8
+IK+RK	VGG-16	46.4	43.2	60.6	25.8	41.2	40.4	30.7	38.8	40.9
Oracle	VGG-16	**49.6**	**44.8**	**67.6**	**28.7**	**48.9**	**35.5**	**30.8**	**36.9**	**42.9**
Source-only	ResNet-101	33.5	34.5	37.2	19.2	27.1	6.4	22.8	28.9	26.2
+IK	ResNet-101	36.8	38.5	44.4	16.2	29.8	8.0	21.9	32.2	28.5
+RK	ResNet-101	43.1	42.7	56.6	32.0	38.1	41.0	29.3	37.8	40.1
+IK+RK	ResNet-101	43.0	42.7	60.0	32.3	46.6	38.4	31.2	38.2	41.5
Oracle	ResNet-101	**47.2**	**46.6**	**66.5**	**30.3**	**52.6**	**35.4**	**32.2**	**36.7**	**43.4**

Table 3. Ablation experiments on adaption from Cityscapes to Foggy Cityscapes. IK represents that mining intrinsic attributes knowledge via domain-invariant classifiers. RK denotes that maintaining the consistency of class relational knowledge in different domains.

still exceeds the state of the art [20] by 4.0% mAP. The increase of detection accuracy demonstrate that the model can better adapt to different weather conditions.

Synthetic to Real. Due to the huge cost of manual annotation, training with synthetic data has attracted more and more attention. The second domain transfer scenario is from synthetic images to real ones. As shown in the left part of Table 2, KTNet performs better than the previous methods using gradient reverse layers, which achieves 50.7% AP with a gain of 1.7% over the second-best model.

Cross Camera Adaptation. It is widely existed in the field of autonomous driving that the setup of vehicle-mounted cameras and the layout of street scenes are both different. Another domain-shift task is cross-camera adaptation from KITTI dataset to Cityscapes dataset. The results of different methods are reported in the right part of Table 2. It is obvious that KTNet has reached a higher level (+2.4% AP) than the state of the arts, which consistently validates that knowledge transferring does reduce the domain gap between different scenarios.

Experiments on the above three benchmark settings show that the proposed method yields certain improvements over existing state-of-the-art methods, and the detection accuracy is increased by 1.7% to 2.4%. Next, we will investigate the effectiveness of each designed knowledge transfer modules.

4.4. Ablation Study

In order to further analyze the proposed method, we provide some in-depth ablation studies that are recorded in Table 3. The base model is referred as the source-only trained detector. First, we alternately apply domain-invariant classifiers (IK) and relational consistency constraints (RK) to quantify the benefits brought by each knowledge transfer module. Second, IK and RK are introduced in the base model simultaneously to verify their collaboration capabilities. Finally, to further investigate the robustness of knowledge transfer paradigm, we repeat the above experiments based on a new backbone network (ResNet-101).

With the help of domain-invariant classifier, the accuracy of source-only model has been upgraded, which proves that exploiting common knowledge of foregrounds can boost the transferability of the detector. However, training domain-invariant classifiers can only implicitly align foreground and background features in different domains, and thus the improvements are limited (no more than 2.5%). The task of object detection not only has to distinguish objects and non-objects, but also needs to accurately identify specific categories of the foregrounds. To this end, we design another module to capture class-aware information.

As listed in the third row of Table 3, adding category consistency constraints considerably improves the results, *e.g.*, 19.7% gains compared with the baseline without adaptation. In contrast to domain-invariant classifiers, explicitly constraining the category relationship to be consistent can more directly teach the model to extract class-specific features and alleviate the impact of domain distribution changes. Although relational consistency constraints provide more significant improvements, RK and IK essentially concentrate on different object knowledge. The former considers correlation between various categories, while the latter focuses on extracting common knowledge of all foregrounds in a class-agnostic manner. As such, these two transfer modules act as a different role. Combining both of them is complementary and boosts the highest detection accuracy. The 4-th row of Table 3 also convey the same argument that common attribute knowledge and relational knowledge can cooperate well to improve the performance of detectors in cross-domain scenarios. After replacing VGG-16 with ResNet-101 as the backbone network, it can be observed that compared with non-adaptive models, the detection accuracy of our method is increased consistently. The mean average precision is enhanced by 22.8%/15.3% with VGG-16/ResNet-101 respectively. Oracle indicates the detector is trained and tested on the target domain. An interesting result is worth noting. The performance of KTNet with IK

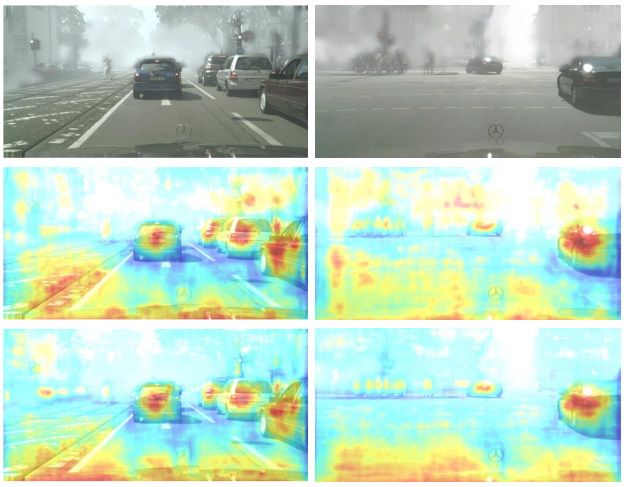

Figure 3. Comparison of object activation response. The first row shows original testing images in the target domain. The second and third rows present attentive regions from source-only model and our KTNet, respectively. Best viewed in color.

and RK in person, rider, truck and mbike is close to that of the oracle model. For train and bike categories, the accuracy of KTNet is even higher.

In summary, we can directly perceive the superiority of our proposed knowledge transfer modules. Comprehensive experiments have confirmed the argument that "focusing on object-related and domain-independent knowledge can effectively improve the robustness and generalization of the detector". The results also demonstrate that our designs can be generalized to different feature extractors.

4.5. Qualitative visual analysis

In addition to quantitative comparison of experimental results, we also provide qualitative visual analysis to illustrate our claim that the transfer of object-related knowledge is conducive to training a domain adaptive detector.

We first exhibit some visualization examples that our model tries to localize foreground objects. In Figure 3, the source-only model pays more attention to the background, but after introducing the knowledge transfer modules, KTNet is able to maintain activation response to foreground regions and suppress the focus on irrelevant background noise. The warmer the color, the stronger the response.

Thanks to effective knowledge transferring, the detection network can activate objects of interest more accurately in the target domain and thus leads to better adaption results. As displayed in Figure 4, green, red and blue boxes are true positives, false positives and false negatives. Due to domain discrepancy, the non-adaptive models (1st row) only respond to some salient objects. Missing detection is an urgent problem to be solved. On the other hand, the proposed method (2nd row) can not only detect more foregrounds (i.e., increase the true positives), but also reduce false alarms. Even when the source and target domains have completely different styles, our model can still localize and identity objects correctly.

Figure 4. Qualitative examples on the target domain. The results of the first and second rows are from source-only model and KTNet. The first column to the second represent two cross-domain settings: Cityscapes→Foggy Cityscapes, SIM10k→Cityscapes. More visualizations are available in the supplementary material.

5. Conclusion

In this paper, we have proposed a cross-domain detection model based on the idea of knowledge transferring, which can be regarded as a new paradigm of domain adaptive framework, in addition to adversarial training and generative models. Specifically, we present two novel knowledge transfer modules that can be applied as plug-and-play components. First, the domain-invariant classifier is designed to explore common attribute knowledge of foreground objects in source and target domains, which effectively aligns feature distributions from another perspective. Second, we model the relationship between object categories, and maintain this inherent knowledge by explicitly constraining the consistency of category correlation in different domains. The incorporation of these two delicately designed modules further refreshes the best historical performance under various cross-domain settings. Our study also reveals a crucial aspect to the success of adaptive detection, that is, focusing on object-related and domain-invariant knowledge can effectively improve the robustness of detectors in different testing scenarios.

Acknowledgment

This research was supported by the National Key Research and Development Program of China under Grant No. 2018AAA0100400, and the National Natural Science Foundation of China under Grants 91646207, 62076242, 62071466, and 61976208.

References

[1] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *ICCV*, pages 2425–2433, 2015.

[2] Tadas Baltrusaitis, Amir Zadeh, Yao Chong Lim, and Louis-Philippe Morency. Openface 2.0: Facial behavior analysis toolkit. In *FG*, pages 59–66. IEEE, 2018.

[3] Shai Ben-David, John Blitzer, Koby Crammer, Alex Kulesza, Fernando Pereira, and Jennifer Wortman Vaughan. A theory of learning from different domains. *Machine learning*, 79(1):151–175, 2010.

[4] Philipp Bergmann, Tim Meinhardt, and Laura Leal-Taixe. Tracking without bells and whistles. In *ICCV*, pages 941–951, 2019.

[5] Zhaowei Cai and Nuno Vasconcelos. Cascade r-cnn: Delving into high quality object detection. In *CVPR*, pages 6154–6162, 2018.

[6] Chaoqi Chen, Weiping Xie, Wenbing Huang, Yu Rong, Xinghao Ding, Yue Huang, Tingyang Xu, and Junzhou Huang. Progressive feature alignment for unsupervised domain adaptation. In *CVPR*, pages 627–636, 2019.

[7] Long Chen, Xin Yan, Jun Xiao, Hanwang Zhang, Shiliang Pu, and Yueting Zhuang. Counterfactual samples synthesizing for robust visual question answering. In *CVPR*, pages 10800–10809, 2020.

[8] Yuhua Chen, Wen Li, Christos Sakaridis, Dengxin Dai, and Luc Van Gool. Domain adaptive faster r-cnn for object detection in the wild. In *CVPR*, pages 3339–3348, 2018.

[9] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *CVPR*, pages 3213–3223, 2016.

[10] Piotr Dollár, Ron Appel, Serge Belongie, and Pietro Perona. Fast feature pyramids for object detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(8):1532–1545, 2014.

[11] Piotr Dollár, Zhuowen Tu, Pietro Perona, and Serge Belongie. Integral channel features. In *BMVC*, 2009.

[12] Pedro F Felzenszwalb, Ross B Girshick, and David McAllester. Cascade object detection with deformable part models. In *CVPR*, pages 2241–2248. IEEE, 2010.

[13] Yaroslav Ganin and Victor Lempitsky. Unsupervised domain adaptation by backpropagation. In *ICML*, pages 1180–1189. PMLR, 2015.

[14] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *CVPR*, pages 3354–3361. IEEE, 2012.

[15] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016.

[16] Zhenwei He and Lei Zhang. Multi-adversarial faster-rcnn for unrestricted object detection. In *ICCV*, pages 6668–6677, 2019.

[17] Zhenwei He and Lei Zhang. Domain adaptive object detection via asymmetric tri-way faster-rcnn. *arXiv preprint arXiv:2007.01571*, 2020.

[18] Judy Hoffman, Eric Tzeng, Taesung Park, Jun-Yan Zhu, Phillip Isola, Kate Saenko, Alexei Efros, and Trevor Darrell. Cycada: Cycle-consistent adversarial domain adaptation. In *ICML*, pages 1989–1998. PMLR, 2018.

[19] Cheng-Chun Hsu, Yi-Hsuan Tsai, Yen-Yu Lin, and Ming-Hsuan Yang. Every pixel matters: Center-aware feature alignment for domain adaptive object detector. In *ECCV*, pages 733–748. Springer, 2020.

[20] Han-Kai Hsu, Chun-Han Yao, Yi-Hsuan Tsai, Wei-Chih Hung, Hung-Yu Tseng, Maneesh Singh, and Ming-Hsuan Yang. Progressive domain adaptation for object detection. In *WACV*, pages 749–757, 2020.

[21] Matthew Johnson-Roberson, Charles Barto, Rounak Mehta, Sharath Nittur Sridhar, Karl Rosaen, and Ram Vasudevan. Driving in the matrix: Can virtual worlds replace human-generated annotations for real world tasks? In *ICRA*, pages 746–753. IEEE, 2017.

[22] Guoliang Kang, Lu Jiang, Yi Yang, and Alexander G Hauptmann. Contrastive adaptation network for unsupervised domain adaptation. In *CVPR*, pages 4893–4902, 2019.

[23] Taekyung Kim, Minki Jeong, Seunghyeon Kim, Seokeon Choi, and Changick Kim. Diversify and match: A domain adaptive representation learning paradigm for object detection. In *CVPR*, pages 12456–12465, 2019.

[24] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *CVPR*, pages 2117–2125, 2017.

[25] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *ICCV*, pages 2980–2988, 2017.

[26] Yong-Xiang Lin, Daniel Stanley Tan, Wen-Huang Cheng, and Kai-Lung Hua. Adapting semantic segmentation of urban scenes via mask-aware gated discriminator. In *ICME*, pages 218–223. IEEE, 2019.

[27] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. Ssd: Single shot multibox detector. In *ECCV*, pages 21–37. Springer, 2016.

[28] Mingsheng Long, Zhangjie Cao, Jianmin Wang, and Michael I Jordan. Conditional adversarial domain adaptation. In *NeurIPS*, pages 1647–1657, 2018.

[29] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *NeurIPS*, 32:8026–8037, 2019.

[30] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *CVPR*, pages 779–788, 2016.

[31] Joseph Redmon and Ali Farhadi. Yolo9000: better, faster, stronger. In *CVPR*, pages 7263–7271, 2017.

[32] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *NeurIPS*, pages 91–99, 2015.

[33] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 115(3):211–252, 2015.

[34] Kuniaki Saito, Yoshitaka Ushiku, Tatsuya Harada, and Kate Saenko. Strong-weak distribution alignment for adaptive object detection. In *CVPR*, pages 6956–6965, 2019.

[35] Kuniaki Saito, Kohei Watanabe, Yoshitaka Ushiku, and Tatsuya Harada. Maximum classifier discrepancy for unsupervised domain adaptation. In *CVPR*, pages 3723–3732, 2018.

[36] Christos Sakaridis, Dengxin Dai, and Luc Van Gool. Semantic foggy scene understanding with synthetic data. *International Journal of Computer Vision*, 126(9):973–992, 2018.

[37] Swami Sankaranarayanan, Yogesh Balaji, Carlos D Castillo, and Rama Chellappa. Generate to adapt: Aligning domains using generative adversarial networks. In *CVPR*, pages 8503–8512, 2018.

[38] Jian Shen, Yanru Qu, Weinan Zhang, and Yong Yu. Wasserstein distance guided representation learning for domain adaptation. In *AAAI*, volume 32, 2018.

[39] Kevin J Shih, Saurabh Singh, and Derek Hoiem. Where to look: Focus regions for visual question answering. In *CVPR*, pages 4613–4621, 2016.

[40] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[41] Sayanan Sivaraman and Mohan Manubhai Trivedi. Looking at vehicles on the road: A survey of vision-based vehicle detection, tracking, and behavior analysis. *IEEE Transactions on Intelligent Transportation Systems*, 14(4):1773–1795, 2013.

[42] Mingxing Tan, Ruoming Pang, and Quoc V Le. Efficientdet: Scalable and efficient object detection. In *CVPR*, pages 10781–10790, 2020.

[43] Zhi Tian, Chunhua Shen, Hao Chen, and Tong He. Fcos: Fully convolutional one-stage object detection. In *ICCV*, pages 9627–9636, 2019.

[44] Eric Tzeng, Judy Hoffman, Kate Saenko, and Trevor Darrell. Adversarial discriminative domain adaptation. In *CVPR*, pages 7167–7176, 2017.

[45] Eric Tzeng, Judy Hoffman, Ning Zhang, Kate Saenko, and Trevor Darrell. Deep domain confusion: Maximizing for domain invariance. *arXiv preprint arXiv:1412.3474*, 2014.

[46] Jasper Uijlings, K E Sande, Theo Gevers, and A W M Smeulders. Selective search for object recognition. *International Journal of Computer Vision*, 104(2):154–171, 2013.

[47] Nicolai Wojke, Alex Bewley, and Dietrich Paulus. Simple online and realtime tracking with a deep association metric. In *ICIP*, pages 3645–3649. IEEE, 2017.

[48] Rongchang Xie, Fei Yu, Jiachao Wang, Yizhou Wang, and Li Zhang. Multi-level domain adaptive learning for cross-domain detection. In *ICCVW*, pages 0–0, 2019.

[49] Chang-Dong Xu, Xing-Ran Zhao, Xin Jin, and Xiu-Shen Wei. Exploring categorical regularization for domain adaptive object detection. In *CVPR*, pages 11724–11733, 2020.

[50] Minghao Xu, Jian Zhang, Bingbing Ni, Teng Li, Chengjie Wang, Qi Tian, and Wenjun Zhang. Adversarial domain adaptation with domain mixup. In *AAAI*, volume 34, pages 6502–6509, 2020.

[51] Jiahui Yu, Yuning Jiang, Zhangyang Wang, Zhimin Cao, and Thomas Huang. Unitbox: An advanced object detection network. In *ACM MM*, pages 516–520, 2016.

[52] Yangtao Zheng, Di Huang, Songtao Liu, and Yunhong Wang. Cross-domain object detection through coarse-to-fine feature adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13766–13775, 2020.

[53] Xingyi Zhou, Vladlen Koltun, and Philipp Krähenbühl. Tracking objects as points. In *ECCV*, pages 474–490. Springer, 2020.

[54] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *ICCV*, pages 2223–2232, 2017.

[55] Xinge Zhu, Jiangmiao Pang, Ceyuan Yang, Jianping Shi, and Dahua Lin. Adapting object detectors via selective cross-domain alignment. In *CVPR*, pages 687–696, 2019.