Cross-Domain Detection Transformer Based on Spatial-Aware and Semantic-Aware Token Alignment

Jinhong Deng, Xiaoyue Zhang, Wen Li, Member, IEEE, Lixin Duan, and Dong Xu, Fellow, IEEE

Abstract—Detection transformers such as DETR (Carion et al., 2020) have recently exhibited promising performance for many object detection tasks, but the generalization ability of those methods is still quite limited for cross-domain adaptation scenarios. To address the cross-domain issue, a straightforward method is to perform token alignment with adversarial training in transformers. However, its performance is often unsatisfactory because the tokens in detection transformers are quite diverse and represent different spatial and semantic information. In this paper, we propose a new method for cross-domain detection transformers called spatial-aware and semantic-aware token alignment (SSTA). Specifically, we take advantage of the characteristics of cross-attention as used in the detection transformer and propose spatial-aware token alignment (SpATA) and semantic-aware token alignment (SemTA) strategies to guide the token alignment across domains. For spatial-aware token alignment, we extract the information from the cross-attention map (CAM) to align the distribution of tokens according to their attention to object queries. For semantic-aware token alignment, we inject the category information into the cross-attention map and construct domain embedding to guide the learning of a multi-class discriminator to model the category relationship and achieve category-level token alignment during the entire adaptation process. We conduct extensive experiments on several widely-used benchmarks, and the results clearly show the effectiveness of our proposed approach over existing state-of-the-art methods.

Index Terms—Detection transformer, domain adaptation, object detection.

I. INTRODUCTION

OBJECT detection [1], [2], [3], [4] is a fundamental problem for visual understanding, which plays an essential role in many multimedia tasks, including image captioning [5], [6], video understanding [7], [8], [9], [10], scene graph generalization [11], etc. With the thriving of deep convolutional neural networks (CNN) [12], [13], many CNN-based object detection approaches (e.g., Faster RCNN [2] and FCOS [3]) have been proposed in the last decade. Recently, detection transformers (e.g., DETR [1]) have gained increasing attention from researchers. Based on the design of visual transformers, detection transformers remove the requirement of hand-designed components of traditional CNN-based object detection methods such as non-maximum suppression (NMS) and anchor generation and simultaneously achieve new state-of-the-art performance in many object detection tasks [1], [14], [15], [16], [17], [18]. Despite the success of detection transformers, the cross-domain generalization ability remains a challenge when adapting a learned model to a novel domain (i.e., target domain). Usually, existing detection transformers suffer from severe performance degradation due to domain discrepancy between the source and target domains [19].

However, addressing the domain shift issue for detection transformers is nontrivial. Researchers have proposed many ways to improve the cross-domain generalization ability of CNN-based object detectors. For example, a variety of studies for cross-domain object detection (CDOD) [20], [21], [22], [23], [24] have been proposed to eliminate the domain discrepancy by aligning the feature distributions of the source and target via adversarial training [20], [21], [25], [26]. Similarly, for the cross-domain detection transformer, a potential and straightforward solution for the cross-domain detection transformer is to perform token alignment with adversarial training, since the visual features are often converted into tokens as the input to the transformer blocks. However, aligning the token distributions is difficult, especially when there is a significant domain gap between domains.

Recent work [19] has attempted to apply adversarial training strategies to tokens in transformers, but the improvements are still unsatisfactory. One of the major reasons is that tokens in detection transformers are quite diverse. In detection transformers (e.g., DETR), the tokens are passed through several multi-head self-attention layers to obtain new token embeddings for representing different spatial and semantic information. Then, object queries are introduced to probe useful tokens and leverage those tokens to predict the positions and categories of different objects. On the one hand, since some tokens are more useful while less for others, it is desirable to take the importance of tokens into consideration in the cross-domain detection.
transformer. On the other hand, the semantic information embedded in tokens is also helpful for aligning token distributions w.r.t. the corresponding category, which can ease the adversarial training process.

In this work, we propose a new cross-domain detection method named spatial-aware and semantic-aware token alignment (SSTA) under the transformer framework. Specifically, we take advantage of the characteristics of cross-attention as used in the detection transformers and two newly developed strategies, i.e., spatial-aware token alignment (SpaTA) and semantic-aware token alignment (SemTA), to guide the token alignment across domains. The cross-attention in the decoder of SSTA utilizes the object queries to aggregate information from encoder outputs (tokens). During this process, only a small part of them are attended to for detecting objects accurately. For spatial-aware token alignment, we can extract the information from the cross-attention map (CAM) to align the distribution of tokens according to their attention to object queries. For semantic-aware token alignment, we inject the category information into the cross-attention map and construct domain embeddings to guide the learning of a multi-class domain discriminator to model the category relationship and achieve category-level alignment during the entire adaptation process.

We conduct extensive experiments on three domain adaptive benchmarks, including adverse weather, synthetic-to-real, and scene adaptation, and achieved new state-of-the-art performance for cross-domain object detection. The experimental results show the effectiveness of our proposed method. We also demonstrated the usefulness of each component in our approach by conducting careful ablation studies. The contributions of our work are three-fold:

- We propose a novel approach for cross-domain object detection under the transformer framework named spatial-aware and semantic-aware token alignment (SSTA). To the best of our knowledge, this is the first attempt to explore the intrinsic cross-attention property for improving the cross-domain generalization ability of detection transformers.
- Two new modules, i.e., spatial-aware token alignment (SpaTA) and semantic-aware token alignment (SemTA), are developed respectively to align the token distributions according to their attention to object queries and to achieve the category-level alignment.
- We conduct extensive experiments on several widely-used benchmarks (e.g., FoggyCityscapes, Sim10K, and BDD100K), and promising results demonstrate the effectiveness of our proposed method over existing state-of-the-art baselines.

II. RELATED WORK

A. Object Detection

Object detection aims to recognize and localize one or multiple objects in a given image. Traditional object detection methods [2], [3], [27], [28], [29] are based on convolutional neural networks (CNN) [12], [13], [30] and can be divided into two directions: one-stage and two-stage methods. One-stage methods [3], [29] directly predict the category and coordinates of objects while two-stage methods [2], [27], [28] first generate some region proposals and then refine their prediction. Although these CNN-based detectors have achieved remarkable breakthroughs, they require many hand-designed components, such as removing duplicated detections by non-maximum suppression and anchor generation. Recently, Carion et al., proposed DETR [1], which achieves end-to-end object detection without anchor generation and any sophisticated post-process procedure. Many DETR-like models [14], [15], [16], [17] have been proposed to further improve the performance of the DETR model in both convergence speed and accuracy. For example, Deformable DETR [14] adopts a deformable attention mechanism [31] into DETR and designs a multi-scale attention module so that it reduces the training time and improves detection performance significantly. Nevertheless, these methods suffer from severe performance degradation due to the domain discrepancy between the training and test domains. To address this problem, we present spatial-aware and semantic-aware token alignment (SSTA) to learn domain-invariant token representations. Following SFA [19], we chose Deformable DETR [14] as the base detector for a fair comparison.

B. Cross-Domain Object Detection

Cross-domain object detection (CDOD) aims to transfer detection knowledge from the label-rich source domain to the label-scarce target domain by bridging the domain discrepancy between them. Many domain adaptation methods [32], [33] have been proposed to address domain shift for the classification task. For object detection, previous works [20], [21], [22], [23], [24], [34], [35], [36], [37], [38], [39] can be roughly categorized into image translation, domain alignment, and self-supervision. Image translation methods [34], [35] and domain alignment methods [20], [21], [40] eliminate the domain discrepancy at the pixel-level and feature-level [20], [21], [40], [41], respectively. One of the promising directions for domain alignment is adversarial training, which has been widely used for CDOD. For example, MSA [41] proposes to align the distribution of multi-scale features with robust discriminative learning. Beyond adversarial training, many metric learning-based methods [38], [40], [42] have been proposed and minimize certain distance metrics between the source and target domains. GPA [40] seeks category-level domain alignment via elaborate prototype representations with graph learning. Self-supervision approaches [22], [23], [37], [43], [44] deploy the pseudo-labeling techniques to provide additional supervision signals for the target domain. AT [44] leverages the adversarial training to improve the pseudo-label quality for better self-supervision on the target domain.

Recently, TIA [45] proposes a new alignment mechanism in classification and regression task spaces instead of feature spaces. MGA [46] designs a unified multi-granularity alignment framework towards domain-invariant feature learning including pixel-, instance-, and category-levels. SIGMA [47] reformulates the domain adaptation with graph matching and completes the semantics during matching. DAOD-AdFW [48]
introduces a new adversarial gradient reversal layer to perform adversarial mining for the hard examples together and generate an auxiliary domain by data augmentation to enforce a new domain-level metric regularization. AIR [49] proposes Adversarial Image Reconstruction as the regularizer to facilitate the adversarial training of the feature extractor. The connection between our proposed method and these advanced works is that adversarial training is used to align the distribution between the source and target domains.

The difference resides primarily in the following aspects. Firstly, these methods are designed for Faster RCNN or FCOS, our method instead addresses the transferability of cross-domain detection transformer. Secondly, our method addresses domain shift by utilizing both spatial and semantic information to aid in token alignment. Our approach is orthogonal to those used in previous works, and we believe that the techniques proposed in these recent studies can be useful in the development of our approach.

Beyond Faster RCNN or FCOS, the transferability of detection transformers remains a challenge. SFA [19] developed a domain adaptive detection transformer to align domain query features and token-wise features and designed an additional bipartite matching consistency loss to enhance the feature discriminability. CCFA [50] aligns the distribution of features with rich context. O^2 net [51] aligns the foreground region identified by the pseudo-label in the backbone outputs and utilizes sliced Wasserstein distance to reduce the domain shift on decoder features. MTTrans [52] follows the paradigm of self-training and is based on the Mean Teacher framework to produce pseudo-label. DA-DETR [4] uses a simple domain discriminator to align the hard-align features obtained by a hybrid attention module. Different from existing works, our SSTA takes advantage of the cross-attention map and leverages the spatial and semantic information to help the token distribution alignment.

### III. METHODOLOGY

In the task of CDOD, we are given a source domain consisting of labeled images with object bounding boxes and their class labels and a target domain consisting of unlabeled images. Let us denote $D_s = \{(x^s_i, y^s_i)\}_{i=1}^{N_s}$ drawn from distribution $P_s$ as the labeled source domain and $D_t = \{(x^t_j)\}_{j=1}^{N_t}$ drawn from distribution $P_t$ as the unlabeled target domain, where $P_s \neq P_t$. And $y^t_j = \{(b^t_j, c^t_j)\}_{j=1}^{N_t}$, where $b^t_j \in \mathbb{R}^4$ and $c^t_j \in \{1, \ldots, C\}$ are the bounding box and corresponding category for each object, respectively, and $m$ is the total number of objects in an image $x^s_i$. Our goal is to learn an object detection model that performs well in the target domain.

#### A. Motivation

In this section, we provide a brief preliminary description of the DETR model. Then, we demonstrate the cross-domain challenges in DETR as well as our new solution.

**DEtection TRansformer (DETR):** DETR consists of a CNN backbone, transformer encoder and transformer decoder. The image $x \in \mathbb{R}^{3 \times H \times W}$ is firstly fed into the CNN backbone (e.g., ResNet50 [13]) to generate a lower-resolution feature map $f \in \mathbb{R}^{C \times H \times W}$, where $C = 2048$, $H = \frac{H}{2}$ and $W = \frac{W}{2}$. The encoder uses a $1 \times 1$ convolution to reduce the channel dimension $d$ and then collapses the spatial dimensions into one dimension, resulting in token inputs $z_c \in \mathbb{R}^{d \times N_k}$, where $N_k = WH$ is the length of the sequence. The encoder layer adopts tokens $z_c$, along with position embedding to interact among tokens and outputs new tokens $z_e \in \mathbb{R}^{d \times N_k}$ through standard architecture that consists of a multi-head self-attention and a feed-forward network (FFN). The decoder comprises multi-head self-attention and multi-head cross-attention mechanisms. Different from the encoder, the decoder first deploys self-attention for $N_q$ object queries and then uses cross-attention (i.e., encoder-decoder attention) to aggregate features from the outputs of the encoder, resulting in a sequence $z_d \in \mathbb{R}^{d \times N_k}$. Finally, the decoder will result in $N_q$ predictions. DETR utilizes the Hungarian algorithm to find bipartite matching between the sets of predictions and ground truth. The loss of DETR can be summarized as follows:

$$L_{det} = L_{cls} + L_{reg},$$

where $L_{cls}$ is for classification and $L_{reg}$ is for bounding box regression. DETR requires much longer training epochs (i.e., 500) to converge than traditional detectors and has relatively low detection accuracy on small objects. Thus Deformable DETR [14] adopts an efficient deformable attention module to replace the dense attention in DETR. The deformable attention mechanism can be naturally extended to aggregate multi-scale features, leading to fast convergence and high performance. Following [19], we chose Deformable DETR [14] as the base detector for a fair comparison.

**Cross-Domain Challenges in DETR:** To improve the generalization ability of detection transformer, a potential solution is to perform token alignment with adversarial learning. Recent work [19] has also attempted to apply adversarial training strategies on tokens in transformers, but the improvements are still unsatisfactory. One of the main reasons is that the tokens in detection transformer are quite diverse. In detection transformers (e.g., DETR), the tokens are passed through several multi-head self-attention layers to obtain new token embeddings for representing different spatial and semantic information. Then, object queries are introduced to probe useful tokens and leverage those tokens to predict the positions and categories of different objects. On the one hand, since some tokens are more useful while less for others, it is desirable to take the importance of tokens into consideration in the cross-domain detection transformer. On the other hand, the semantic information embedded in tokens is also helpful for aligning the token distributions of the corresponding category. This would ease the adversarial training when aligning the token distributions between domains.

To this end, we propose the spatial-aware token alignment (SpaTA) and semantic-aware token alignment (SemTA) strategies to guide the token alignment across domains by leveraging the characteristics of cross-attention in the detection transformer. As shown in Fig. 2, the proposed spatial-aware and the semantic-aware token alignment (SSTA) module adopts the cross-attention map (CAM) and predictions of the decoder to
align the distributions of tokens from the CNN and encoder. The details will be presented below.

B. Vanilla Token Alignment

Before we provide the details of the design of our SSTA module, we first introduce vanilla token alignment. The existing adversarial methods [20], [21], [53] usually take a discriminator to reduce domain discrepancy by aligning feature distributions between domains. The discriminator tries to distinguish which domain the features come from, while the feature extractor aims to confuse features and deceive the discriminator in a minimax manner. It can be placed at a certain layer or multiple layers of the feature extractor. In practice, a gradient reverse layer (GRL) [25] is used to connect the discriminator and feature extractor and flips the gradients when it flows through the feature extractor.

To bridge the domain gap, a naive solution is to simply align the distribution of tokens where the domain discriminator tries to recognize each token. As shown in Fig. 1, we feed the source images into the detection model (i.e., DETR) and calculate the detection loss $L_{det}$ in (1) with the supervision of labels in the source domain. The source and target tokens from CNN or transformer encoder are fed into the discriminator through a GRL (Gradient Reversal Layer) to calculate the vanilla token alignment loss $L_{ta}$ in (2).

![Fig. 1. Illustration of vanilla token alignment. The source images are fed into the detection model (i.e., DETR) and calculate the detection loss $L_{det}$ in (1) with the supervision of labels in the source domain. The source and target tokens from CNN or transformer encoder are fed into the discriminator through a GRL (Gradient Reversal Layer) to calculate the vanilla token alignment loss $L_{ta}$ in (2).](image)

where $\lambda$ is the trade-off parameter, and $L_{ta}^c$ and $L_{ta}^e$ are the vanilla token alignment loss for the CNN and encoder tokens.

C. Spatial-Aware Token Alignment

As in the analysis in Section III-A, object queries are introduced to probe useful tokens and leverage those tokens to predict the positions and categories of different objects. In other words, tokens contribute differently to the detection results. Simply aligning the token distribution between domains leads to unsatisfactory improvements, as tokens in the detection transformer have different importance to the object detection task. If we consider the tokens equally contributing to the adversarial training, we will overlook matching the distribution of critical tokens that may contain essential instances and global context for accurately predicting the positions and categories of different objects. Consequently, efforts to reduce the domain gap will eventually encounter difficulties, making the alignment less effective.

Motivated by this, we propose a spatial-aware token alignment (SpaTA) module to discover instance-related tokens and emphasize their alignment by assigning higher weights to these tokens for adversarial training according to their attention to the object queries. Formally, we can obtain the objective as follows:

$$L_{spa} = \sum_{i=1}^{N_q} (1 + W^i) \cdot L_{ta}^i,$$

where $W^i$ is the weight for the $i$-th token; intuitively, the more important the token should be assigned higher weights. As shown in the bottom part of Fig. 2, we utilize the cross-attention map (CAM) as an alternative to providing the weights, as object queries probe features by giving different weights to tokens via the cross-attention mechanism.

For dense attention, we can easily obtain the cross-attention map (CAM) in the decoder by averaging attention maps from every decoder layer. However, for the case of deformable attention, the CAM cannot be directly obtained in deformable attention because of its special design, i.e., the spatial distribution of attention in deformable DETR is not uniform. Concretely, to obtain the CAM in deformable attention, we scatter and accumulate the cross-attention in the decoder from each object query to discrete token positions in the sequence. In Deformable DETR [14], each query usually predicts a reference point, four offsets, and its corresponding offsets, the attention weights, and values, respectively. Thus, the values are calculated in cross-attention as follows:

$$\tilde{v} = \sum_{i} A \cdot B(t_i, r + \Delta r) \cdot v(t_i),$$

where $B(\cdot, \cdot)$ is the bilinear interpolation operation kernel, $t_i$ enumerates all integral spatial locations in the feature map and $v(t_i)$ is the value at location $t_i$. 

Fig. 2. **Top:** The overview of our method. We design a new spatial-aware and semantic-aware token alignment (SSTA) module to align CNN token and encoder token distributions across two domains. We take advantage of the characteristics of the cross-attention in the decoder and feed the cross-attention map (CAM) and the predictions of the detection head (FFN) to improve the token alignment. **Bottom:** Overview of our spatial-aware and semantic-aware token alignment (SSTA) module. Consider the SSTA module for encoder tokens as an example. The proposed SSTA module takes the tokens as the input and jointly utilizes semantic-aware token alignment (SemTA) and spatial-aware token alignment (SpaTA) to respectively align token distributions. SemTA affiliates the predictions of the detection head into the cross-attention map (CAM) and obtains a category cross-attention map (CCAM), which can be used to construct domain embedding to guide the learning of a multi-class discriminator (MCD) to achieve category-level token alignment. The SpaTA utilizes the CAM to give different weights to the adversarial learning of tokens according to their attention to object queries.

The values are obtained by aggregating values from surrounding tokens at \( t_i \). According to (5), for the attention to each token, we can obtain the CAM of the \( i \)-th query by averaging the cross-attention map over all decoder layers as follows:

\[
\mathcal{M}_i = \frac{1}{N_d} \sum_{l=1}^{N_d} \sum_{(A_l, r, \Delta r)} A_l \cdot B(t, r + \Delta r),
\]

where \( N_d \) is the number of decoder layers.

After obtaining the CAM, we filter out some attention that is less than a given threshold. In summary, the important weight for tokens can be obtained via:

\[
W = \mathcal{M} \odot \mathbb{1}(\mathcal{M} \geq \tau(\mathcal{M})),
\]

where \( \mathcal{M} \) is the average of CAM for all queries and \( \tau(\mathcal{M}) = \text{mean}(\mathcal{M}) \) is an adaptive threshold for each sample \( x \). \( \mathbb{1} \) is the indicator function.

### D. Semantic-Aware Token Alignment

Although we have discovered the critical tokens to emphasize their alignment and avoid the influence of noise tokens, the model still has the risk of misalignment during the adaptation process [36], [60]. The semantic information of tokens is helpful for aligning the token distributions of the corresponding category so that the model can avoid class misalignment. For example, the “car” and the “truck” instances are forced to be very close in the feature space, deteriorating the model discriminant.
ability. Therefore, we propose to utilize a multi-class discriminator [60] (MCD) to capture the category information during adversarial training so that it realizes category-level token alignment. The multi-class discriminator contains not only domain information but also category relationships. Concretely, we remold the single-class discriminator to a multi-classes discriminator that outputs 2K logits, where \( K = C + 1 \), \( K \) for the source domain, and others for the target domain. The domain embeddings \( d \in \mathbb{R}^{2K \times 1} \) of the source and target are \([0; s]\) and \([s; 0]\), respectively, where \( s \in \mathbb{R}^{K \times 1} \) is the domain knowledge and \( 0 \in \mathbb{R}^{K \times 1} \) is the all-zero vector. The objective of semantic-aware token alignment can be written as follows:

\[
\mathcal{L}_{sem}^{i} = -\sum_{k=1}^{2K} d_k \cdot \log(\hat{D}(z_i)_k),
\]

(8)

where \( \hat{D} \) is the multi-class domain discriminator. The key factor is determining how to obtain the domain knowledge \( s \) to build domain embedding for these tokens. As illustrated in the bottom part of Fig. 2, we also utilize CAM to extract domain knowledge by injecting the category information into it. Specifically, we affilitate the predictions of the detection head into the CAM and obtain a category cross-attention map (CCAM), which can be formally defined as follows:

\[
\hat{M}_k = \frac{1}{N_q^k} \sum_{i=1}^{N_q^k} 1(\hat{y}_i = k) \cdot M_i,
\]

(9)

where \( \hat{M}_k \in \mathbb{R}^{N_q} \) refers to CCAM \( \hat{M} \in \mathbb{R}^{N_q \times K} \) for category \( k \) and \( N_q^k \) is the number of queries that belong to category \( k \). \( \hat{y}_i \) is the category prediction from the detection head for the \( i \)-th query and \( 1(\cdot) \) is the indicator function where if \( \cdot \) is true, then it equals 1; otherwise it equals 0. \( s \) can be obtained after applying the softmax function to the CCAM \( \hat{M} \). Finally, we obtain our domain adaptation loss by replacing \( \mathcal{L}_{da}^{i} \) with the semantic-aware token alignment \( \mathcal{L}_{sem}^{i} \) in (4):

\[
\mathcal{L}_{da} = \sum_{i=1}^{N_q} (1 + \mathcal{W}^i) \cdot \mathcal{L}_{sem}^{i},
\]

(10)

E. Overall Objective

In summary, the overall objective includes the detection loss of Deformable DETR [14] on the source domain and domain adaptation loss for the CNN and encoder tokens. Thus, the overall objective can be defined as:

\[
\mathcal{L} = \mathcal{L}_{det} + \lambda \cdot (\mathcal{L}_{da}^{d} + \mathcal{L}_{da}^{a}),
\]

(11)

where \( \lambda \) is the trade-off parameter, \( \mathcal{L}_{da}^{d} \) and \( \mathcal{L}_{da}^{a} \) are the domain adaptation losses for the CNN and encoder tokens, respectively.

IV. EXPERIMENTS

Following [19], we train the model with labeled source data and unlabeled target data and test on the target data. We conduct extensive experiments on three CDOD scenarios. The detection results are evaluated with mean Average Precision (mAP) under the threshold of 0.5.

A. Datasets

Cityscapes: Cityscapes dataset is collected for the scenes understanding of road and street. It comprises 2,975 and 500 images for training and validation, respectively. It contains 8 categories: person, rider, car, truck, bus, train, motorcycle, and bicycle.

FoggyCityscapes: FoggyCityscapes [65] dataset is the foggy version of Cityscapes and generated using the depth information provided by Cityscapes. Thus, it shares the common annotations with Cityscapes. It contains three levels for foggy weather, including 0.01, 0.15, and 0.02. In experiments, we choose the worst foggy weather (i.e., 0.02).

Sim10k: Sim10K [66] dataset is a synthetic dataset rendered by the gaming engine Grand Theft Auto V (GTAV). This dataset contains 10,000 images with 58,701 bounding boxes with the category of “car”.

DD100k: BDD100K [67] dataset is a large-scale autonomous driving and contains 100 k images with six types of weather, six different scenes, and three categories for the time of day. We extract the subset of daytime, resulting in 36,728 training and 5,258 validation images.

Following existing works [20], [53], we evaluate our method on three benchmark settings:

- **Weather Adaptation**: We take Cityscapes as the source domain and FoggyCityscape as the target domain, and the model is trained on the train set of Cityscapes and FoggyCityscape and evaluated on the validation split of FoggyCityscapes.
- **Syn2Real**: We explore the adaptation of Sim10K to Cityscapes, we train the model using all the images of Sim10K and the train split of Cityscapes, and report mAP on the validation split of Cityscapes with “car” category.
- **Scene Adaptation**: We use Cityscapes as the source domain dataset and BDD100K containing distinct scenes as a large unlabeled target domain dataset. We evaluate the model on the validation set of BDD100K.

B. Implementation Details

Following the default setting in SFA [19], we adopt Deformable DETR [14] as the base detector, which contains the ResNet-50 [13] backbone pretrained on ImageNet [68], six transformer encoders, six transformer decoders and multiple prediction heads. We adopt the Adam [69] optimizer to update parameters. For Cityscapes to FoggyCityscapes, we first train the model with a learning rate 2 \( \times 10^{-4} \) for 40 epochs and then decay the learning rate to 2 \( \times 10^{-5} \) for 10 more epochs. The trade-off parameter \( \lambda \) is set to 1.0. For Sim10K to Cityscapes and Cityscapes to BDD100K, we set the initial learning rate and the trade-off parameter \( \lambda \) to 5 \( \times 10^{-5} \) and 0.01 respectively. We pre-train models on source data to obtain reliable CAM. All the experiments were conducted using four V100 GPUs with a batch size of 16, i.e., each GPU contains 2 source images and 2 target images. We implemented our method with the PyTorch deep learning framework.
TABLE I

AVERAGE PRECISIONS (%) OF DIFFERENT METHODS ON CITYSCAPES → FOGGY CITYSCAPES

| Method          | Detector | person | rider | car   | truck | bus   | train | mcycle | bicycle | mAP  |
|-----------------|----------|--------|-------|-------|-------|-------|-------|--------|---------|------|
| Faster RCNN     |          | 26.9   | 38.2  | 35.6  | 18.3  | 32.4  | 9.6   | 25.8   | 28.6    | 26.9 |
| DA-Faster [20]  |          | 25.0   | 31.0  | 40.5  | 22.1  | 35.3  | 20.2  | 20.0   | 27.1    | 27.6 |
| SWDA [21]       |          | 29.9   | 42.3  | 43.5  | 24.5  | 36.2  | 32.6  | 30.0   | 35.3    | 34.3 |
| CFDA [55]       |          | 43.2   | 37.4  | 52.1  | 34.7  | 34.0  | 46.9  | 29.9   | 30.8    | 38.6 |
| UMT [22]        |          | 33.0   | 46.7  | 48.6  | 34.1  | 56.5  | 46.8  | 30.4   | 37.4    | 41.7 |
| MeGA [36]       |          | 37.7   | 49.0  | 52.4  | 25.4  | 49.2  | 46.9  | 34.5   | 39.0    | 41.8 |
| ICCR-VDD [56]   | Faster RCNN | 33.4   | 44.0  | 51.7  | 33.9  | 52.0  | 34.7  | 34.2   | 36.8    | 40.0 |
| VSI-GA [57]     |          | 38.8   | 45.9  | 57.2  | 29.9  | 50.2  | 51.9  | 31.9   | 40.9    | 43.3 |
| DIDN++ [54]     |          | 38.3   | 44.4  | 51.8  | 28.7  | 53.3  | 34.7  | 32.4   | 40.4    | 40.5 |
| TIA [45]        |          | 34.8   | 46.3  | 49.7  | 31.1  | 52.1  | 48.6  | 37.7   | 38.1    | 42.3 |
| DDF [39]        |          | 37.6   | 45.5  | 56.1  | 30.7  | 50.4  | 47.0  | 31.1   | 39.8    | 42.3 |
| MGA [46]        |          | 43.9   | 49.6  | 60.6  | 29.6  | 50.7  | 39.0  | 38.3   | 42.8    | 44.3 |
| DAO-D-ADF [48]  |          | 36.5   | 46.7  | 54.3  | 30.3  | 51.2  | 48.7  | 31.6   | 39.1    | 42.3 |
| FCOS [3]        | Faster RCNN | 36.9   | 36.3  | 44.1  | 18.6  | 29.3  | 8.4   | 20.3   | 31.9    | 28.2 |
| EPM [58]        |          | 41.9   | 38.7  | 56.7  | 22.6  | 41.5  | 26.8  | 24.6   | 35.5    | 36.0 |
| SCAN [42]       |          | 41.7   | 43.9  | 57.3  | 28.7  | 48.6  | 48.7  | 31.0   | 37.3    | 42.1 |
| SCAN++ [38]     |          | 44.2   | 43.9  | 57.9  | 28.2  | 48.1  | 51.2  | 30.1   | 39.5    | 42.8 |
| KTN [59]        |          | 46.4   | 43.2  | 60.6  | 25.8  | 41.2  | 40.4  | 30.7   | 38.8    | 40.9 |
| SS [23]         |          | 45.1   | 47.4  | 59.4  | 24.3  | 50.0  | 25.7  | 26.0   | 38.7    | 39.6 |
| MGA [46]        |          | 43.1   | 47.3  | 61.5  | 30.2  | 53.2  | 50.3  | 27.9   | 36.9    | 43.8 |
| SIGMA [47]      |          | 44.0   | 43.9  | 60.3  | 31.6  | 50.4  | 51.5  | 31.7   | 40.6    | 44.2 |
| AIR [49]        |          | 43.6   | 46.7  | 62.1  | 27.8  | 44.0  | 37.0  | 29.9   | 38.4    | 41.2 |
| DefDETR [14]    | DefDETR  | 38.6   | 40.6  | 45.8  | 11.6  | 28.9  | 1.7   | 18.9   | 39.1    | 28.1 |
| SPA [19]        |          | 46.5   | 48.6  | 62.6  | 25.1  | 46.2  | 29.4  | 28.3   | 44.0    | 41.3 |
| CCFA [50]†      |          | 48.2   | 47.3  | 63.9  | 23.7  | 47.8  | 39.0  | 33.6   | 42.8    | 43.3 |
| DA-DETR ‡       |          | 49.9   | 50.0  | 63.1  | 24.0  | 45.8  | 37.5  | 31.6   | 46.3    | 43.5 |
| SSTA (Ours)     |          | 50.5   | 53.0  | 67.2  | 24.7  | 47.7  | 33.0  | 36.7   | 46.6    | 44.9 |

DeFDET is the abbreviation for deformable DETR. † and ‡ denote iterative bounding box refinement and two-stage Deformable DETR. Note that we denote the domain adaptation extension in DIDN [54] as DIDN++.

C. Results

We have conducted extensive experiments and validated the effectiveness of our method by comparing various state-of-the-art CDOD methods, mainly including three kinds of methods: 1) two-stage detector Faster RCNN, 2) one-stage detector FCOS and 3) Deformable DETR. For all the methods, we report the results from the original papers. To validate the effectiveness of our proposed method, we also report the results of the Source model where the model is only trained on the source domain and evaluated on the target domain.

Weather Adaptation (Cityscapes → FoggyCityscapes): The adaptation results are shown in Table I. We can observe that our proposed method outperforms the previous state-of-the-art approaches by a large margin, reaching 44.9% mAP. Specifically, Deformable DETR (Source) achieves 28.1% mAP, demonstrating that Deformable DETR has decent generalization but still suffers from the distribution discrepancy across domains. Both SFA [19] and our SSTA improve the Source baseline. However, our SSTA improves 3.6% mAP compared with the counterpart SFA [19]. In addition, our SSTA outperforms CCFA [50] and DA-DETR [53] that uses a more advanced variant of Deformable DETR than ours. These results demonstrate that our method by leveraging intrinsic cross-attention to conduct spatial-aware and semantic-aware token alignment can effectively improve the generalization ability of DETR on the target domain.

Syn2Real (Sim10K → Cityscapes): The results are presented in Table II. Our SSTA reaches the highest AP on car (57.7%) that exceeds all compared state-of-the-art methods, including the two-stage, one-stage, and DETR works, by a large margin, that is 3.1% mAP and 2.4% over best-performing one-stage detector MGA [46] and DETR counterpart DA-DETR [53]. These results verify the effectiveness of our SSTA.

Scene Adaptation (Cityscapes → BDD100K): The quantitative results are shown in Table III. According to Table III, our SSTA method achieves the new state-of-the-art results of 29.5% in terms of mAP, which surpasses the previous works. This again demonstrates the generalization of our approach.

D. Analysis

Ablation Studies: To further verify the effectiveness of our proposed method, we have conducted detailed ablation studies by isolating each component of our SSTA. The experimental results are shown in Table IV. In particular, our SpaTA significantly boosts the baseline, leading to 14.4% mAP improvement compared with the Source model (28.1%). This implies that the CAM can provide sufficient information to discover critical tokens, and emphasizing their contributions to distribution alignment will significantly improve the generalization ability of Deformable DETR. Moreover, SemTA also improves the accuracy of Deformable DETR, achieving 43.9% in terms of mAP. These improvements are mainly due to our SemTA considering...
category information during token alignment and thus avoiding class misalignment. By synergizing SpaTA and SemTA together, we obtain 44.9% mAP, which shows they are complementary to each other.

Variants of Deformable DETR: Deformable DETR has been improved in two variants: iterative bounding box refinement (IBR) and two-stage Deformable DETR. We also explore the results of our method on these two variants. From Table V, the results show that our method improves the generalization of the DETR with different variants and suppresses the existing works CCFA [50] and DA-DETR [53] with the corresponding variant by a large margin, respectively.

Parameter Analysis: We also investigate the influence of the trade-off parameter $\lambda$ which is used to balance the weight between the source detection loss $L_{\text{det}}$ and the domain adaptation loss. Table VI summarizes the results on Cityscapes $\rightarrow$ FoggyCityscapes. Note that when $\lambda = 0$, the method degenerates to the Source model. According to Table VI, we can conclude that our SSTA consistently improves the generalization ability of the Deformable DETR in a wide range of $\lambda$, and $\lambda = 1.0$ and $\lambda = 1.5$ are the best among them.

Computational Complexity: Although we introduce adversarial training and multi-class discriminators, compared to the detection transformer model (e.g., the baseline model Deformable DETR), the extra computational complexity of our method is negligible. The multi-class discriminator only contains three fully connected (FC) layers followed by ReLu and Dropout layers. Specifically, our method only introduces extra 2.48 G FLOPs compared with 173 G FLOPs of Deformable DETR during model training. Kindly note that the discriminators are only used for the model training and do not introduce extra overload in the inference stage.

E. Qualitative Results

Cross-Attention Map Visualization: We visualize the cross-attention map (CAM) of the Source model and our SSTA on the Cityscapes to FoggyCityscapes in Fig. 3. We can observe that the attention map obtained by the Source model (top row) generally focuses on the object-related region. Although there also exist certain noisy regions, they can be used as a good starting point to improve adversarial training for domain adaptation. Moreover, in the training process, as the model is trained to adapt to the target domain, the CAM obtained with the new model is improved (bottom row). For example, in the middle image on the top row, the CAM obtained by the Source model misses some objects and also gives biased attention to the left or right boundary of images, while the CAM obtained by the new model is refined. This clearly validates our motivation that the CAM can be used to effectively improve the model training for domain adaptive object detection.

Detection Visualization: As shown in Fig. 4, we provide some detection results of state-of-the-art SFA [19], our SSTA, and ground truth (GT) on the Cityscapes to FoggyCityscapes adaptation. It can be seen that our proposed method outperforms SFA and produces more accurate detection results. As can be seen from Fig. 4, the baseline method (SFA) fails to detect some objects (hidden in the foggy) that are correctly detected by our proposed SSTA approach, e.g., the car in 1st and 2nd rows. Additionally, our proposed SSTA method also proves effective in detecting the minor classes, as evidenced by the successful detection of the rider and bicycle in the 3rd row of the images.

Qualitative Ablation Study: We conduct a qualitative analysis of our proposed components to study the individual impact of each component. According to Fig. 5, we have the following observations. 1) Compared with the Source Only (Fig. 5(a)), the Vanilla TA (Fig. 5(b)) can correctly detect many objects in the foggy by aligning the token distributions. Our SpaTA (Fig. 5(c)) could further improve the Vanilla TA by assigning more weights to the critical instance-related tokens, e.g., the more correctly detected occluded car in 1st row and person in 2nd rows of images. 2) For the component of semantic-aware token alignment (SemTA), the comparison between Fig. 5(b) and (d) demonstrates that SemTA not only detects more common objects (e.g., car) but also eliminate the missing error on rare categories, e.g., correctly detect the bus in 1st and 2nd rows. These results imply that our SemTA helps to achieve better classification by category-level token alignment, avoiding the risk of class misalignment. 3) By combining SpaTA and SemTA, we can observe that our full model SSTA (Fig. 5(e)) could address the above corner cases, showing the complementary effect of the proposed two components.

F. Limitations and Future Works

We conduct a detailed analysis of the limitations of the proposed method by examining the failure case illustrated in Fig. 6. The analysis reveals that the model fails to detect some objects,
TABLE III
AVERAGE PRECISIONS (%) OF DIFFERENT METHODS ON CITYSCAPES → BDD100K

| Methods      | Detector | person | rider | car   | truck | bus   | mcycle | bicycle | mAP   |
|--------------|----------|--------|-------|-------|-------|-------|--------|---------|-------|
| Faster R-CNN (Source) | Faster R-CNN | 28.8   | 25.4  | 44.1  | 17.1  | 16.1  | 13.9   | 22.4    | 24.1  |
| DA-Faster [20] | Faster R-CNN | 28.9   | 27.4  | 44.2  | 19.1  | 18.0  | 14.2   | 22.4    | 24.9  |
| SWDA [21]     | Faster R-CNN | 29.5   | 29.9  | 44.8  | 20.2  | 20.7  | 15.2   | 23.1    | 26.2  |
| SCDA [24]     | Faster R-CNN | 29.3   | 29.2  | 44.4  | 20.3  | 19.6  | 14.8   | 23.2    | 25.8  |
| ECR [64]      | Faster R-CNN | 32.8   | 29.3  | 45.8  | 22.7  | 20.6  | 14.9   | 25.5    | 27.4  |
| FCOS [3] (Source) | FCOS     | 38.6   | 24.8  | 54.5  | 17.2  | 16.3  | 15.0   | 18.3    | 26.4  |
| EPM [58]      | FCOS     | 39.6   | 26.8  | 55.8  | 18.8  | 19.1  | 14.5   | 20.1    | 27.8  |
| DefDETR [14] (Source) | DefDETR | 38.4   | 27.1  | 56.1  | 14.6  | 12.3  | 16.3   | 20.7    | 26.5  |
| CCFA [50]†     | DefDETR | 42.2   | 28.6  | 60.2  | 18.4  | 21.1  | 16.9   | 19.4    | 29.5  |
| SFA [19]      | DefDETR | 40.2   | 27.6  | 57.5  | 19.1  | 23.4  | 15.4   | 19.2    | 28.9  |
| SSTA (Ours)   | DefDETR | 39.4   | 31.9  | 59.4  | 16.3  | 17.7  | 15.3   | 26.2    | 29.5  |

DefDETR is the abbreviation for DeformableDETR. † denotes iterative bounding box refinement.

Fig. 3. Visualization of Cross-attention Map (CAM) on FoggyCityscapes. Top: Source model. Bottom: SSTA (Ours).

Fig. 4. Visualization of detection results on the Weather Adaptation scenario. We use the threshold 0.8 for better illustration. Best appreciated when viewed in color and zoomed up.
DENG et al.: CROSS-DOMAIN DETECTION TRANSFORMER BASED ON SPATIAL-AWARE AND SEMANTIC-AWARE TOKEN ALIGNMENT 5243

Fig. 5. Qualitative comparison results on (a) Source Only: The model is only trained on the source domain. (b) Vanilla token alignment (TA). (c) w/SpaTA: using only the spatial-aware token alignment (SpaTA). (d) w/SemTA: using only the semantic-aware token alignment (SemTA). (e) SSTA: our full model spatial-aware and semantic-aware token alignment (SSTA). (f) Ground Truth labels. (Zooming in for a better view).

TABLE IV
ABLATION STUDIES OF SSTA ON CITYSCAPES → FOGGYCITYSCAPES

| Method          | SpaTA | SemTA | mAP (%) |
|-----------------|-------|-------|---------|
| DeDetR(Source)  | -     | -     | 28.1    |
| Vanilla TA     | -     | -     | 41.3    |
| Proposed       | ✓     | ✓     | 42.5    |
|                | ℹ️    |       | 43.9    |
|                | ✓     | ✓     | 44.9    |

TA indicates token alignment.

TABLE V
RESULTS ON DIFFERENT VARIANTS OF DEFORMABLE DETR

| Method        | IBR   | Two-stage | CS→FCS | CS→BDD |
|---------------|-------|-----------|--------|--------|
| CCF2          | ✓     |           | 43.3   | 29.5   |
| DA-Detr       | ✓     | ✓         | 43.5   | -      |
| SSTA          | ✓     | ✓         | 44.9   | 29.5   |
|               | ✓     | ✓         | 46.4   | 31.2   |
|               | ✓     | ✓         | 47.2   | 33.6   |

IBR means iterative bounding box refinement. Two-stage stands for two-stage Deformable DETR. CS→FCS and CS→BDD denote Cityscapes → FoggyCityscapes and Cityscapes → BDD100K, respectively.

TABLE VI
AVERAGE PRECISIONS (%) W.R.T. DIFFERENT VALUES OF λ, WHICH IS USED TO BALANCE THE WEIGHT BETWEEN THE SOURCE DETECTION LOSS AND THE DOMAIN ADAPTATION LOSS ON CITYSCAPES → FOGGYCITYSCAPES

| λ   | 0.0 | 0.1 | 0.5 | 1.0 | 1.5 | 2.0 |
|-----|-----|-----|-----|-----|-----|-----|
| SSTA| 28.1| 42.1| 44.3| 44.9| 44.9| 44.6|

Fig. 6. Illustration of failure cases of the proposed SSTA compared with the Ground Truth label.

particularly small-scale ones, leading to poor-quality representation. To address this issue, we can leverage multi-scale feature fusion and pyramid to enhance the feature representation. Moreover, it is also possible to integrate scale information of objects into the adversarial training to achieve scale-aware token alignment.

V. CONCLUSION

Detection transformers (e.g., DETR) have shown promising results for object detection, when training and test images come from the same domain. However, they usually do not work well for cross-domain problems. In this work, we address cross-domain object detection by proposing a novel approach named semantic-aware and spatial-aware token alignment (SSTA) under the transformer framework. In SSTA, two new modules i.e., spatial-aware token alignment (SpaTA) and semantic-aware token alignment (SemTA), are developed to guide token alignment across domains. Promising results on benchmark datasets demonstrate the effectiveness of our method.

REFERENCES

[1] N. Carion et al., “End-to-end object detection with transformers,” in Proc. Eur. Conf. Comput. Vis., 2020, pp. 213–229.
[2] S. Ren, K. He, R. Girshick, and J. Sun, “Faster R-CNN: Towards real-time object detection with region proposal networks,” in Proc. Adv. Neural Inf. Process. Syst., 2015, pp. 91–99.
[3] Z. Tian, C. Shen, H. Chen, and T. He, “FCOS: Fully convolutional one-stage object detection,” in Proc. IEEE/CVF Int. Conf. Comput. Vis., 2019, pp. 9626–9635.
[4] C. Chen and Q. Ling, “Adaptive convolution for object detection,” IEEE Trans. Multimedia, vol. 21, pp. 3205–3217, 2019.
[5] M. Yang et al., “Multitask learning for cross-domain image captioning,” IEEE Trans. Multimedia, vol. 21, pp. 1047–1061, 2019.
[6] J. H. Tan, C. S. Chan, and J. H. Chuah, “COMIC: Toward a compact image captioning model with attention,” IEEE Trans. Multimedia, vol. 21, pp. 2686–2696, 2019.
[7] A. B. Jelodar, D. Paulius, and Y. Sun, “Long activity video understanding using functional object-oriented network,” IEEE Trans. Multimedia, vol. 21, pp. 1813–1824, 2019.
[8] M. Li et al., “Video pivoting unsupervised multi-modal machine translation,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 45, no. 3, pp. 3918–3932, Mar. 2023.
[9] J. Deng et al., “Single shot video object detector,” IEEE Trans. Multimedia, vol. 23, pp. 846–858, 2021.
[10] L. Zhang et al., “TN-ZSTAD: Transferable network for zero-shot temporal activity detection,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 45, no. 3, pp. 3848–3861, Mar. 2023.
[11] X. Chang et al., “A comprehensive survey of scene graphs: Generation and application,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 45, no. 1, pp. 1–26, Jan. 2023.
[12] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “Imagenet classification with deep convolutional neural networks,” in Proc. Adv. Neural Inf. Process. Syst., 2012, pp. 1097–1105.
[13] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2016, pp. 770–778.

[14] X. Zhu et al., “Deformable DETR: Deformable transformers for end-to-end object detection,” in Proc. Int. Conf. Learn. Representations, 2020.

[15] B. Roh, J. Shin, W. Shin, and S. Kim, “Sparse DETR: Efficient end-to-end object detection with learnable sparsity,” in Proc. Int. Conf. Learn. Representations, 2022.

[16] M. Xu, H. Wang, B. Ni, Q. Tian, and W. Zhang, “Cross-domain detection via graph-induced prototype alignment,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2020, pp. 12355–12364.

[17] Y. Pan, A. J. Ma, Y. Gao, J. Wang, and Y. Lin, “Multi-scale adversarial cross-domain detection with robust discriminative learning,” in Proc. Winter Conf. Appl. Comput. Vis., 2020, pp. 1324–1332.

[18] W. Li, X. Liu, X. Yao, and Y. Yuan, “Scan: Cross domain object detection with semantic conditioned adaptation,” in Proc. AAAI Conf. Artif. Intell., 2022, pp. 1421–1428.

[19] A. RoyChoudhury et al., “Automatic adaptation of object detectors to new domains using self-training,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2019, pp. 880–889.

[20] Y.-J. Li et al., “Cross-domain adaptive teacher for object detection,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2022, pp. 7581–7590.

[21] L. Zhao and L. Wang, “Task-specific inconsistency alignment for domain adaptive object detection,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., 2022, pp. 14217–14226.

[22] W. Zhou, D. Du, L. Zhang, T. Luo, and Y. Wu, “Multi-granularity alignment domain adaptation for object detection,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., 2022, pp. 9581–9590.

[23] W. Li, X. Liu, and Y. Yuan, “SIGMA: Semantic-complete graph matching for domain adaptive object detection,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2022, pp. 5291–5300.

[24] J. Li et al., “Domain adaptive object detection for autonomous driving under foggy weather,” in Proc. Winter Conf. Appl. Comput. Vis., 2023, pp. 612–622.

[25] K. Sun et al., “AIR-DA: Adversarial image reconstruction for unsupervised domain adaptive object detection,” IEEE Robot. Autom. Lett., vol. 8, no. 6, pp. 3645–3652, Jun. 2023.

[26] S. Chen, “Conditional context-aware feature alignment for domain adaptive detection transformer,” in Proc. Int. Conf. Multimedia Model., 2022, pp. 272–283.

[27] K. Gong et al., “Improving transferability for domain adaptive detection transformers,” in Proc. ACM Int. Conf. Multimedia, 2022, pp. 1543–1551.

[28] J. Yu et al., “Cross-domain object detection with mean-teacher transformer,” in Proc. Eur. Conf. Comput. Vis., 2022, pp. 629–645.

[29] J. Zhang, J. Huang, Z. Luo, G. Zhang, and S. Lu, “DA-DETR: Domain adaptive detection transformer by hybrid attention,” 2021, arXiv:2103.17084.

[30] C. Lin et al., “Domain-invariant disentangled network for generalizable object detection,” in Proc. IEEE/CVF Int. Conf. Comput. Vis., 2021, pp. 8751–8760.

[31] Y. Zheng, D. Huang, S. Liu, and Y. Wang, “Cross-domain object detection through coarse-to-fine feature adaptation,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2020, pp. 13766–13775.

[32] A. Wu, R. Liu, Y. Han, L. Zhu, and Y. Yang, “Vector-decomposed disentanglement for domain-invariant object detection,” in Proc. IEEE/CVF Int. Conf. Comput. Vis., 2021, pp. 9342–9351.

[33] F. Rezaeianaran et al., “Seeking similarities over differences: Similarity-based domain alignment for adaptive object detection,” in Proc. IEEE/CVF Int. Conf. Comput. Vis., 2021, pp. 9204–9213.

[34] C.-C. Hsu, Y.-H. Tsai, Y.-Y. Lin, and M.-H. Yang, “Every pixel matters: Center-aware feature alignment for domain adaptive object detector,” in Proc. 6th Eur. Conf. Comput. Vis., 2020, pp. 733–748.

[35] K. Tian, C. Zhang, Y. Wang, S. Xiang, and C. Pan, “Knowledge mining and transferring for domain adaptive object detection,” in Proc. IEEE/CVF Int. Conf. Comput. Vis., 2021, pp. 9133–9142.

[36] H. Wang, T. Shen, W. Zhang, L.-Y. Duan, and T. Mei, “Classes matter: A fine-grained adversarial approach to cross-domain semantic segmentation,” in Proc. Eur. Conf. Comput. Vis., 2020, pp. 8869–8878.

[37] Z. He and L. Zhang, “Multi-adversarial faster-RCNN for unrestricted object detection,” in Proc. IEEE/CVF Int. Conf. Comput. Vis., 2019, pp. 6668–6677.

[38] 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., 2020, pp. 11724–11733.

[39] C. Li et al., “Spatial attention pyramid network for unsupervised domain adaptation,” in Proc. Winter Conf. Appl. Comput. Vis., 2020, pp. 841–847.

[40] C.-D. Xu, X.-R. Zhao, X. Jin, and X.-S. Wei, “Exploring categorical regularization for domain adaptive object detection,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., 2021, pp. 11724–11733.

[41] C. Sakaridis, D. Dai, and L. Van Gool, “Semantic fogy scene understanding with synthetic data,” Int. J. Comput. Vis., vol. 126, no. 9, pp. 973–992, 2018.
[66] M. Johnson-Roberson et al., “Driving in the matrix: Can virtual worlds replace human-generated annotations for real world tasks?,” in Proc. IEEE Int. Conf. Robot. Automat., 2016, pp. 746–753.

[67] F. Yu et al., “BDD100K: A diverse driving dataset for heterogeneous multi-task learning,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2020, pp. 2633–2642.

[68] J. Deng et al., “ImageNet: A large-scale hierarchical image database,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2009, pp. 248–255.

[69] D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” in Proc. Int. Conf. Learn. Representations, 2015.

Jinhong Deng received the B.Eng. degree from Southwest Jiaotong University, Chengdu, China, in 2019. He is currently working toward the Ph.D. degree with the School of Computer Science and Engineering, University of Electronic Science and Technology of China, Chengdu, China. His research interests include transfer learning, label-efficient learning, and their applications in computer vision.

Xiaoyue Zhang received the B.S. degree from Nankai University, Tianjin, China, the B.B.A dual degree from Tianjin University, Tianjin, in 2019, and the M.S. degree from the Hong Kong University of Science and Technology, Hong Kong, SAR, China, in 2021. She is currently working toward the Ph.D. degree with the School of Computer Science and Engineering, University of Electronic Science and Technology of China, Chengdu, China. Her research interests include multi-modal affective computing, vision-language learning, and transfer learning.

Wen Li (Member, IEEE) received the B.E. and M.S. degrees from Beijing Normal University, Beijing, China, in 2007 and 2010, respectively, and the Ph.D. degree from Nanyang Technological University, Singapore, in 2015. He is currently a Full Professor with the School of Computer Science and Engineering and Shenzhen Institute for Advanced Study, University of Electronic Science and Technology of China, Shenzhen, China. His research interests include transfer learning, multi-view learning, multiple Kernel learning, and their applications in computer vision.

Lixin Duan received the B.E. degree from the University of Science and Technology of China, Hefei, China, in 2008, and the Ph.D. degree from Nanyang Technological University, Singapore, in 2012. He is currently a Full Professor with the School of Computer Science and Engineering, University of Electronic Science and Technology of China, Chengdu, China. His research interests include machine learning algorithms especially in transfer learning and domain adaptation and their applications in computer vision. He was the recipient of the Microsoft Research Asia Fellowship in 2009 and Best Student Paper Award at IEEE Conference on Computer Vision and Pattern Recognition 2010.

Dong Xu (Fellow, IEEE) received the B.E. and Ph.D. degrees from the University of Science and Technology of China, Hefei, China, in 2001 and 2005, respectively. When he was working toward the Ph.D. degree, he was an intern with Microsoft Research Asia, Beijing, China, and a Research Assistant with the Chinese University of Hong Kong, Shatin, Hong Kong, for more than two years. He was a Postdoctoral Research Scientist with Columbia University, New York, NY, USA, for one year. He was a Faculty Member with Nanyang Technological University, Singapore, and the Chair of computer engineering with the University of Sydney, Camperdown, NSW, Australia. He is currently a Professor with the Department of Computer Science, University of Hong Kong, Hong Kong. His research interests include computer vision, statistical learning, and multimedia content analysis. He was the co-author of a paper that was the recipient of the Best Student Paper Award in IEEE Conference on Computer Vision and Pattern Recognition in 2010 and a paper that won the Prize Paper Award in IEEE TRANSACTIONS ON MULTIMEDIA in 2014.

Authorized licensed use limited to the terms of the applicable license agreement with IEEE. Restrictions apply.