DA-DETR: Domain Adaptive Detection Transformer with Information Fusion

Jingyi Zhang[1,*] Jiaxing Huang[1,*] Zhipeng Luo[1,2] Gongjie Zhang[1,3] Xiaoqin Zhang[4] Shijian Lu[1,†]
[1] S-lab, Nanyang Technological University [2] SenseTime Research
[3] Black Sesame Technologies [4] Wenzhou University

Abstract

The recent detection transformer (DETR) simplifies the object detection pipeline by removing hand-crafted designs and hyperparameters as employed in conventional two-stage object detectors. However, how to leverage the simple yet effective DETR architecture in domain adaptive object detection is largely neglected. Inspired by the unique DETR attention mechanisms, we design DA-DETR, a domain adaptive object detection transformer that introduces information fusion for effective transfer from a labeled source domain to an unlabeled target domain. DA-DETR introduces a novel CNN-Transformer Blender (CTBlender) that fuses the CNN features and Transformer features ingeniously for effective feature alignment and knowledge transfer across domains. Specifically, CTBlender employs the Transformer features to modulate the CNN features across multiple scales where the high-level semantic information and the low-level spatial information are fused for accurate object identification and localization. Extensive experiments show that DA-DETR achieves superior detection performance consistently across multiple widely adopted domain adaptation benchmarks.

1. Introduction

Object detection aims to predict a bounding box and a class label for interested objects in images and it has been a longstanding challenge in the computer vision research. Most existing work adopts a two-stage detection pipeline that involves heuristic anchor designs, complicated post-processing such as non-maximum suppression (NMS), etc. The recent detection transformer (DETR) [5] has attracted increasing attention which greatly simplifies the two-stage detection pipeline by removing hand-crafted anchors [21, 22, 49] and NMS [21, 22, 49]. Despite its great detection performance under a fully supervised setup, how to leverage the simple yet effective DETR architecture in domain adaptive object detection is largely neglected.

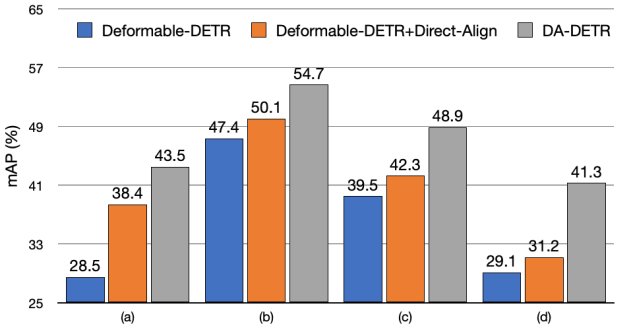

Figure 1. The vanilla *Deformable-DETR* [81] trained with labeled source data cannot handle target data well due to cross-domain shift. The introduction of adversarial feature alignment in *Deformable-DETR + Direct-align* [19] improves the detection clearly. The proposed *DA-DETR* fuses CNN features and transformer features ingeniously which achieves superior unsupervised domain adaptation consistently across four widely adopted benchmarks including Cityscapes → Foggy cityscapes in (a), SIM 10k → Cityscapes in (b), KITTI → Cityscapes in (c) and PASCAL VOC → Clipart1k in (d).

Different from the conventional CNN-based detection architectures such as Faster RCNN [49], DETR has a CNN backbone followed by a transformer head consisting of an encoder-decoder structure. The CNN backbone and the transformer head learn different types of features [17,48,69] - the former largely captures low-level localization features ($e.g.$, edges and lines around object boundaries) while the latter largely captures global inter-pixel relationship and high-level semantic features. At the other end, many prior studies show that fusing different types of features often is often helpful in various visual recognition tasks [9, 11]. Hence, it is very meaningful to investigate how to fuse the two types of DETR features to address the domain adaptive object detection challenge effectively.

We design DA-DETR, a simple yet effective Domain Adaptive DETR that introduces information fusion into the DETR architecture for effective domain adaptive object detection. The core design is a CNN-Transformer Blender (CTBlender) that employs the high-level semantic

*Equal contribution, {jingyi.zhang, jiaxing.huang}@ntu.edu.sg.
†Corresponding author, shijian.lu@ntu.edu.sg.

features in the Transformer head to conditionally modulate the low-level localization features in the CNN backbone. CTBlender consists of two sequential fusion components, including split-merge fusion (SMF) that fuses CNN and Transformer features within an image and scale aggregation fusion (SAF) that fuses the SMF features across multiple feature scales. Different from the existing weight-and-sum fusion [9, 11], SMF first splits CNN features into multiple groups with different semantic information as captured by the Transformer head and then merges them with channel shuffling for effective information communication among different groups. The SMF features of each scale are then aggregated by SAF for fusing both semantic and localization information across multiple feature scales. Hence, CTBlender captures both semantic and localization features ingeniously which enables comprehensive and effective inter-domain feature alignment with a single discriminator.

The main contributions of this work can be summarized in three aspects. *First*, we propose DA-DETR, a simple yet effective domain adaptive detection transformer that introduces information fusion for effective domain adaptive object detection. To the best of our knowledge, this is the first work that explores information fusion for domain adaptive object detection. *Second*, we design a CNN-Transformer Blender that fuses the CNN features and Transformer features ingeniously for effective feature alignment and knowledge transfer across domains. *Third*, extensive experiments show that DA-DETR achieves superior object detection over multiple widely studied domain adaptation benchmarks as compared with the state-of-the-art as shown in Fig. 1.

2. Related Work

Transformers [56] have achieved great success in various neural language processing (NLP) tasks [3, 14, 40, 46, 47] due to their computational efficiency and scalability. Inspired by the success of transformers in NLP, several studies [5, 8, 12, 15, 16, 62, 70, 71, 77, 78, 81] attempt to adapt transformers to computer vision tasks. For example, ViT [16] adopts transformer for image classification, which splits each image into patches and treats them as input tokens to transformers. SETR [78] extends ViT to semantic segmentation by introducing multiple decoders designed for pixel-wise segmentation. For object detection, different from CNN-based detector, DETR [5] treats detection as a set prediction task [50], which eliminates the dependence on various heuristic and hand-crafted designs such as anchor generation, ROI pooling and non-maximum suppression. Existing vision transformers achieve very promising performance in supervised learning. However, how to adapt and generalize them to unsupervised domain adaptation tasks has been largely neglected. In this work, we investigate domain adaptive detection transformers in an unsupervised manner.

Unsupervised Domain Adaptation (UDA) has been studied extensively in recent years, largely for alleviating data annotation constraint in deep network training in various visual recognition tasks [7, 19, 23, 28, 30–32, 41, 42, 44, 52, 58, 63, 66, 73, 74, 82, 83]. For object detection, the target of UDA is to mitigate the domain gap between a source domain and a target domain, so that the source data can be employed to train better detectors for target data. Most existing domain adaptive detectors [6, 7, 36, 51, 55, 57, 64, 76, 79, 80] adopt CNN-based detector ($i.e.$, Faster R-CNN) and achieve UDA via adversarial learning [7, 26, 36, 51, 55, 64, 79], image translation [1, 1, 29, 33, 35, 38, 54, 67] and self-training [30, 67]. However, little research [59, 68] is conducted on how to adopt DETR in domain adaptive detection tasks, $e.g.$, SFA [59] tackles the domain adaptive object detection via query-based feature alignment and token-wise feature alignment. Differently, we introduce the information fusion idea into the DETR architecture for effective domain adaptive object detection.

Feature Fusion is often helpful in various visual recognition tasks. For example, ResNet [25] fuses features of different network layers with skip connections which achieves obvious performance gains along with increased network depth. Feature Pyramid Network (FPN) [39] fuses features of different scales to build high-level semantic feature maps, and it usually improves the object detection with clear margins in various detection tasks. Some work [9, 11] instead achieves feature fusion with channel-wise attention [27, 45] and spatial-wise attention [4, 37, 60, 61]. For example, AFF [11] presents a multi-scale channel attention module for better fusing features from different layers and scales that capture different types of semantics. LS-DeconvNet [9] introduces a gated fusion layer to fuse RGB and depth features effectively. We explore information fusion for domain adaptive detection transformer, and design a CNN-Transformer Blender for fusing CNN features and Transformer features for effective domain adaptive object detection. Different from [9, 11] that adopts a weight-and-sum strategy, our CNN-Transformer Blender employs the Transformer features to modulate the CNN features for more effective adversarial feature alignment.

3. Preliminaries of Detection Transformer

DETR [5] consists of a CNN backbone [25] to extract features, an encoder-decoder transformer and a simple feed forward network (FFN) to make final detection prediction. Given an image x, the CNN backbone G first generates feature f and then reshapes f to a vector. The encoder-decoder in DETR follows the standard architecture of the transformer [56], which consists of multiple multi-head self-

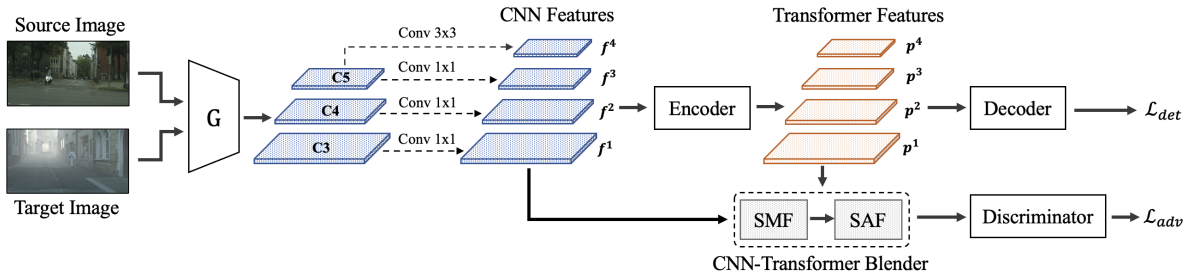

Figure 2. Overview of the proposed DA-DETR: the proposed DA-DETR consists of a base detector (including a backbone G and a transformer encoder-decoder), a discriminator and a CNN-Transformer Blender (CTBlender). Given an input image from either source or target domain, the backbone G first produces multi-scale CNN features f^l ($l = 1, 2, 3, 4$) and then feeds them to the transformer encoder to obtain Transformer features p^l ($l = 1, 2, 3, 4$). For *supervised learning*, the Transformer features generated by the source images are further fed to the decoder to compute supervised detection loss $\mathcal{L}_{det}$ with the corresponding ground truth. For *unsupervised learning*, CTBlender takes f^l and p^l as inputs for feature fusion. Finally, the output of CTBlender is fed to the discriminator for computing an adversarial loss $\mathcal{L}_{adv}$ which drives adversarial alignment of source and target features.

attention modules that are defined by:

$$MSA(z_q, f) = \sum_{h=1}^{H} P_H[\sum_k SA_{hqk} \cdot P_H{}' f_k], \quad (1)$$

where $MSA(\cdot)$ consists of H single attention heads, z_q and f_k denotes representation features of query element and key element, $P_H \in \mathbb{R}^{d \times d_h}$ and $P_H{}' \in \mathbb{R}^{d \times d_h}$ are learnable projection weights ($d_h = d/H$, where d is the dimension of f). Each self-attention weight SA_{hqk} is a type of scaled dot-product attention, which maps a query and a set of key-value pairs into an output:

$$SA_{hqk} \propto \exp\left(\frac{z_q^T U_m^T V_m f_c}{\sqrt{d_h}}\right), \quad (2)$$

where $U_m, V_m \in \mathbb{R}^{d_h \times d}$ are also learnable weights.

We adopt Deformable-DETR [81] as the base detector. Different from the conventional DETR [5], Deformable-DETR replaces the normal attention with the deformable attention which improves the convergence speed greatly:

$$DeformableMSA(z_q, p_q, f) = \sum_{h=1}^{H} P_H[\sum_k SA_{hqk} \cdot P_H{}' f(p_q + \delta p_{hqk})], \quad (3)$$

where δp_{hqk} and SA_{hqk} denote the sampling offset and attention weight of the k-th sampling point in the m-th attention head, respectively. Such sampling design significantly mitigates the slow convergence and high complexity issues of DETR [5]. In addition, Deformable-DETR is extended to aggregating multi-scale features as shown in Fig. 2. The multi-scale feature maps f^l ($l = 1, 2, 3, 4$) are extracted from the output of Block C3-C5 in the ResNet backbone [25]. More specifically, f^1, f^2 and f^3 are extracted from the output feature maps of Block C3-C4 via a 1×1 convolution. The lowest resolution feature map, *i.e.*, f^4, is extracted by 3×3 stride 2 convolution on the output feature maps of Block C5. Such multi-scale design enables attention to capture relationships among different-scale features effectively.

4. Method

4.1. Task Definition

The work focuses on the problem of unsupervised domain adaptation (UDA) in object detection. It involves a source domain $\mathcal{D}_s$ and a target domain $\mathcal{D}_t$, where $\mathcal{D}_s = \{(x_s^i, y_s^i)\}_{i=1}^{N_s}$ is fully labeled, and y_s^i represents the labels of the sample image x_s^i. The goal is to train a detection transformer that well performs on unlabeled target-domain data x_t^i. The baseline model is trained with the labeled source data (*i.e.*, $\mathcal{D}_s$) only:

$$\mathcal{L}_{det} = l(T(G(x_s)), y_s), \quad (4)$$

where G denotes backbone, T denotes transformer encoder-decoder and $l(\cdot)$ denotes the supervised detection loss that consists of a matching cost and a Hungarian loss [5, 81] for object category and object box predictions.

4.2. Framework Overview

As shown in Fig. 2, the proposed DA-DETR consists of a base detector (including a backbone G and a transformer encoder-decoder T), a discriminator C_d and a CNN-Transformer Blender (CTBlender). We adopt the deformable-DETR [81] as the base detector, where G extracts features from the input images and T predicts a set of bounding boxes and pre-defined semantic categories according to the extracted features. CTBlender consists of two sub-modules including a *split-merge fusion (SMF)* and a *scale aggregation fusion (SAF)* as in Fig. 3. Taking the

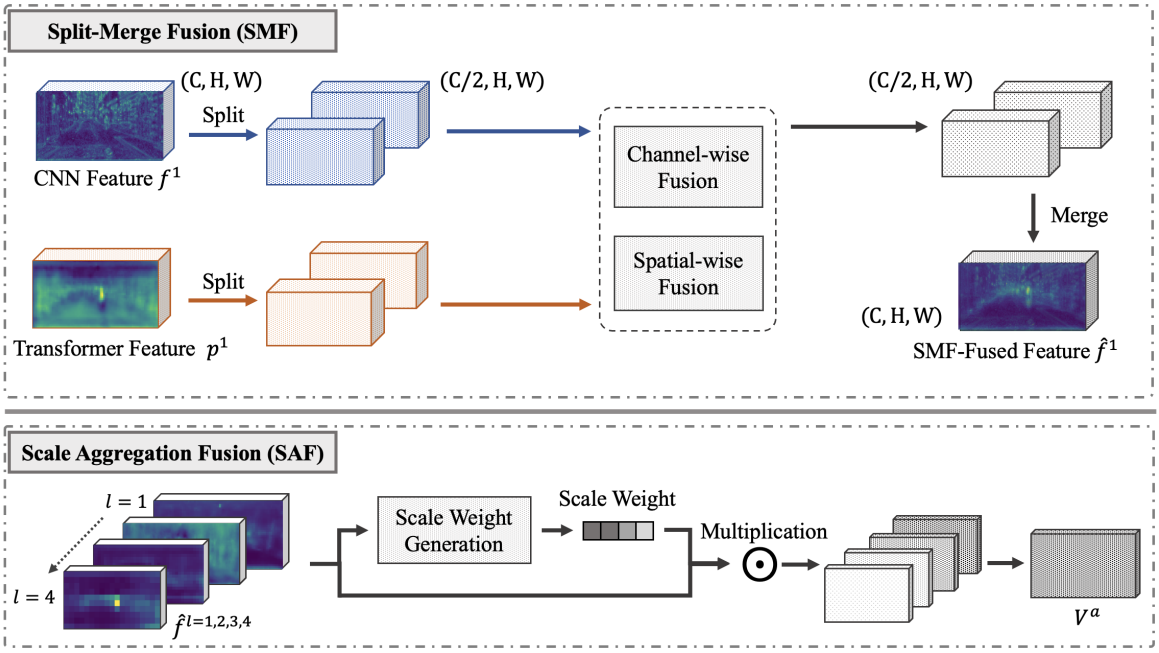

Figure 3. Overview of the proposed CNN-Transformer Blender (CTBlender). CTBlender consists of split-merge fusion (SMF) and scale aggregation fusion (SAF) as illustrated. In SMF, Transformer features of all four scales are adopted to modulate the CNN features. Take the first level f^1 and p^1 as an example. The Transformer feature p^1 and the CNN feature f^1 are divided into K groups ($e.g.$, $K = 2$), and further fed to SMF to perform spatial-wise fusion and channel-wise fusion, respectively. The fused group features are then merged to generate the final fused features $\hat{f}^1$ for the first scale. In SAF, $\hat{f}^l$ ($l = 1, 2, 3, 4$) are aggregated with different scale-wise weights to generate the final feature V^a.

CNN features from G and the Transformer features from the encoder E as inputs, CTBlender fuses the positional and semantic information in Transformer features and the localization information in CNN features for comprehensive and effective feature alignment across domains.

Given an input image from either source or target domain, the backbone G will first produce multi-scale features f^l ($l = 1, 2, 3, 4$) and then feeds them to the transformer encoder to obtain Transformer features p^l ($l = 1, 2, 3, 4$). For *supervised learning*, the Transformer features generated by source images $x_s \in \mathcal{D}_s$ are further fed to decoder to predict a set of bounding boxes and pre-defined semantic categories, which will be used to calculate a detection loss $\mathcal{L}_{det}$ under the supervision of the corresponding ground-truth label $y_s \in \mathcal{D}_s$. For *unsupervised learning*, CTBlender takes f^l ($l = 1, 2, 3, 4$) and p^l ($l = 1, 2, 3, 4$) generated by both source and target images ($i.e.$, $x_s \in \mathcal{D}_s$ and $x_t \in \mathcal{D}_t$) as inputs. Finally, the output of CTBlender is fed to the discriminator C_d to compute an adversarial loss $\mathcal{L}_{adv}$ for inter-domain feature alignment. The overall network is optimized by the adversarial loss $\mathcal{L}_{adv}$ and the detection loss $\mathcal{L}_{det}$.

4.3. CNN-Transformer Blender

One key component in DA-DETR is a CNN-Transformer Blender (CTBlender) that fuses different features for effective domain alignment. CTBlender takes multi-scale CNN features and the corresponding multi-scale Transformer features as the input, where the semantic and positional information in the Transformer features are fused with the localization information in the CNN features via split-merge fusion (SMF). The SMF-fused features are then aggregated across multiple scales via scale aggregation fusion (SAF).

Split-Merge Fusion. In SMF, the rich semantic and positional information in the the multi-scale Transformer features $p = \{p^l\}_{l=1}^{L}$ are exploited to fuse with multi-scale CNN features $f = \{f^l\}_{l=1}^{L}$ for adversarial feature alignment across domains. As SMF operations at each feature scale are the same, we take the first scale $l = 1$ as an example to illustrate how we perform split-merge fusion.

Inspired by the split-fuse-merge in [72], SMF first splits CNN features into multiple groups and then fuses them with the Transformer features. After that, the fused features are merged with channel shuffling for effective information communication among different groups. Given a Transformer feature $p^1 \in \mathbb{R}^{C\times H\times W}$ and a backbone CNN feature $f^1 \in \mathbb{R}^{C\times H\times W}$ (C, H, W indicate the number of channel of feature map, and the height and the width of feature map, respectively), p^1 is first split into K groups evenly along channels, $i.e.$, $\{p_k^1\}_{k=1}^{K} \in \mathbb{R}^{(C/K)\times H\times W}$, where

each group captures different semantic information of the input image. The fusion in each group is then achieved via spatial-wise fusion and channel-wise fusion, respectively.

For the spatial-wise fusion, the split Transformer features are firstly fed into a normalization layer and then re-weighted by a learnable weight map and a learnable bias map:

$$\hat{p}_{ks}^{1} = f_s\left(w_s \cdot GN\left(p_k^1\right) + b_s\right), \tag{5}$$

where $f_s(\cdot)$ is an activation function that limits the input in the range of $[0, 1]$.

For the channel-wise fusion, the split Transformer feature is firstly compacted by the Global Average Pooling (GAP) and then re-weighted by a learnable weight vector and a learnable bias vector:

$$\hat{p}_{kc}^{1} = f_s\left(w_c \cdot GAP(p_k^1) + b_c\right), \tag{6}$$

where $f_s(\cdot)$ is an activation function that limits the input to the range of $[0, 1]$.

Similar to the operation for Transformer feature p^1, the CNN feature f^1 is also divided into K groups along the channels, *i.e.*, $\left\{f_k^1\right\}_{k=1}^{K} \in \mathbb{R}^{(C/K)\times H\times W}$.

We further adopt shuffle operation to enable information communication across channels [43, 75]. Specifically, we first re-weight the split CNN feature by the corresponding re-weighted Transformer feature (*i.e.*, $\hat{p}_{ks}^1$ and $\hat{p}_{kc}^1$) to generate re-weighted split CNN feature $\hat{f}_k^1$.

Then we shuffle $\hat{f}_k^1$ along channels to enable information flow across channels for better feature fusion. Lastly, we conduct the above operations K times to generate K shuffled features for each group, *i.e.*, $\left\{\hat{f}_k^1\right\}_{k=1}^{K} \in \mathbb{R}^{(C/K)\times H\times W}$. The shuffled features are concatenated to obtain the fused feature map $\hat{f}^1 \in \mathbb{R}^{C\times H\times W}$:

$$\hat{f}^1 = f_c\left(\hat{f}_1^1, ..., \hat{f}_k^1, ..., \hat{f}_K^1\right), \tag{7}$$

where similar operations are conducted to get the results of all levels $\hat{f} = \left\{\hat{f}^l\right\}_{l=1}^{L}$.

Scale Aggregation Fusion. To explicitly perform feature fusion of different scales, we design a scale aggregation fusion (SAF) to aggregate features $\hat{f}$ with different scale weights as illustrated in the bottom part of Fig. 3.

Specifically, we compact each scale of feature $\hat{f} = \left\{\hat{f}^l\right\}_{l=1}^{L}$ into a channel-wise vector $u = \left\{u^l\right\}_{l=1}^{L} \in \mathbb{R}^{C\times 1\times 1}$ via a Global Average Pooling (GAP) layer. The scale weights α_l are obtained from channel-wise vectors u^l. Firstly, the channel-wise vectors are merged together to obtain merged vector u_m by an element-wise addition.

Then, a fully connected layer separates u_m to L scale-weight vectors $\alpha^l \in \mathbb{R}^{C\times 1\times 1}$. Finally, V^a is obtained by

Direct-align	+SMF		+SAF	mAP
	Shuffling	Splitting		
				28.5
✓				38.4
✓	✓			41.4
✓		✓		41.8
✓	✓	✓		42.3
✓			✓	41.7
✓	✓	✓	✓	**43.5**

Table 1. Ablation study of DA-DETR over domain adaptation task Cityscapes → Foggy Cityscapes.

$$V^a = \sum_{l=1}^{L} \hat{f}^l \cdot \alpha^l, \tag{8}$$

where V^a is a highly embedded feature that captures rich semantic information and localization information.

4.4. Network Training

The network is trained with two losses, *i.e.*, a supervised object detection loss $\mathcal{L}_{det}$ as defined in Eq. 4 and an adversarial alignment loss $\mathcal{L}_{adv}$ that is defined as follow:

$$\begin{aligned}\mathcal{L}_{adv} = \; & \mathbb{E}_{(f,p)\in\mathcal{D}_s} \log C_d\left(\mathcal{H}\left(f,p\right)\right) \\ & + \mathbb{E}_{(f,p)\in\mathcal{D}_t} \log\left(1 - C_d\left(\mathcal{H}\left(f,p\right)\right)\right),\end{aligned} \tag{9}$$

where $f = G(x)$ and $p = E(G(x))$. G denotes backbone; E denotes transformer encoder; $\mathcal{H}$ denotes CNN-Transformer Blender (CTBlender) and C_d denotes the discriminator. Both source images x_s and target images x_t are utilized to compute adversarial loss.

In summary, the overall optimization objective of DA-DETR is formulated by

$$\max_{C_d} \min_{G,T,\mathcal{H}} \mathcal{L}_{\text{det}}(G,T) - \lambda\mathcal{L}_{\text{adv}}(\mathcal{H}, C_d), \tag{10}$$

where T denotes the transformer in DETR, λ is the weight factor that balances the influences of $\mathcal{L}_{det}$ and $\mathcal{L}_{adv}$ in training. Note that we adopt a gradient reverse layer (GRL) [19] to enable the gradient of $\mathcal{L}_{adv}$ to be reversed before back-propagating to $\mathcal{H}$ from C_d.

5. Experiments

This section presents experimentation including experiment setups, implementation details, ablation studies, comparisons with the state-of-the-art and discussion. More details are to be described in the ensuing subsections.

5.1. Experiment Setups

Datasets. Following [7, 26, 33, 35, 51, 64], we evaluate DA-DETR under four widely adopted domain adaptation scenarios with eight datasets as listed: 1) *Normal Weather to Foggy Weather* (Cityscapes [10] → Foggy

Cityscapes → Foggy cityscapes										
Method	Backbone	person	rider	car	truck	bus	train	mcycle	bicycle	mAP
Deformable-DETR [81]	ResNet-50	37.7	39.1	44.2	17.2	26.8	5.8	21.6	35.5	28.5
DAF [7]	ResNet-50	48.2	48.8	61.5	22.6	43.1	20.2	30.3	42.1	39.6
SWDA [51]	ResNet-50	49.0	49.0	61.4	23.9	43.1	22.9	31.0	45.2	40.7
SCL [55]	ResNet-50	49.4	48.6	61.2	**27.2**	41.1	34.8	28.5	42.5	41.7
GPA [65]	ResNet-50	49.5	46.7	58.6	26.4	42.2	32.3	29.1	41.8	40.8
CRDA [64]	ResNet-50	49.8	48.4	61.9	22.3	40.7	30.0	29.9	45.4	41.1
CF [79]	ResNet-50	49.6	49.7	62.6	23.3	43.4	27.4	30.2	44.8	41.4
SAP [36]	ResNet-50	49.3	49.9	62.5	23.0	44.1	29.4	31.3	45.8	41.9
SFA [59]	ResNet-50	46.5	48.6	62.6	25.1	**46.2**	29.4	28.3	44.0	41.3
MTTrans [68]	ResNet-50	47.7	49.9	65.2	25.8	45.9	33.8	**32.6**	46.5	43.4
DA-DETR	ResNet-50	**49.9**	**50.0**	**63.1**	24.0	45.8	**37.5**	31.6	**46.3**	**43.5**

Table 2. Experimental results (%) of the scenario Normal weather to Foggy weather: Cityscapes → Foggy Cityscapes.

SIM 10k → Cityscapes		
Method	Backbone	mAP on Car
Deformable-DETR [81]	ResNet-50	47.4
DAF [7]	ResNet-50	49.8
SWDA [51]	ResNet-50	50.5
SCL [55]	ResNet-50	51.6
GPA [65]	ResNet-50	51.3
CRDA [64]	ResNet-50	52.1
CF [79]	ResNet-50	52.5
SAP [36]	ResNet-50	52.3
SFA [59]	ResNet-50	52.6
MTTrans [68]	ResNet-50	**57.9**
DA-DETR	ResNet-50	54.7

Table 3. Experimental results (%) of the scenario Synthetic scene to Real scene: SIM 10k → Cityscapes.

KITTI → Cityscapes		
Method	Backbone	mAP on Car
Deformable-DETR [81]	ResNet-50	39.5
DAF [7]	ResNet-50	43.6
SWDA [51]	ResNet-50	44.3
SCL [55]	ResNet-50	44.5
GPA [65]	ResNet-50	43.2
CRDA [64]	ResNet-50	44.8
CF [79]	ResNet-50	45.2
SAP [36]	ResNet-50	46.5
SFA [59]	ResNet-50	46.7
DA-DETR	ResNet-50	**48.9**

Table 4. Experimental results (%) of the scenario Cross-camera Adaptation: KITTI → Cityscapes.

Cityscapes [53]); 2) *Synthetic Scene to Real Scene* (SIM 10k [34] → Cityscapes [10]); 3) *Cross-camera Adaptation* (KITTI [20] → Cityscapes [10]) and 4) *Real-world Images to Artistic Images* (PASCAL VOC [18] → Clipart1k, Watercolor2k, Comic2k [33]).

5.2. Implementation Details

In all experiments, we adopt deformable-DETR [81] as the base detector. Since there is only few prior study [59] on transformer-based domain adaptive detection, we modify existing object detectors using Faster R-CNN [7, 24, 36, 51, 64, 79] to the transformer-based domain adaptive detection for fair comparisons. The modification is accomplished by keeping domain adaptation modules unchanged but replacing post-processing modules in Faster R-CNN ($e.g.$, region proposal network, proposal classification module, etc.) by the encoder-decoder module of deformable DETR. In addition, we adopt ResNet-50 [25] (pre-trained on ImageNet [13]) as backbone, and use SGD optimizer [2] with a momentum 0.9 and a weight decay $1e-4$ in all experiments with deformable-DETR.

In all experiments, the weight factor λ in Eq. 10 is fixed at 0.1 and the number of split groups K in SMF is fixed at 32. All the experiments are implemented in Pytorch. For evaluation metrics, we report average precision (AP) for each object category and mean average precision (mAP) of all object categories with a threshold of intersection over union (IoU) at 0.5 as in [7, 51, 64].

5.3. Ablation Studies

The proposed CTBlender consists of split-merge fusion (SMF) and scale aggregation fusion (SAF). We first study the two fusion modules to examine how they contribute to the overall unsupervised domain adaptive detection performance. Table 1 shows experimental results over the validation data of Foggy Cityscapes under the adaptation scenario 'normal weather to foggy weather'.

As Table 1 shows, the *Baseline* [81] trained using the labeled source data only does not perform well due to domain shifts. The model *Direct-align* aligns CNN features directly via adversarial learning which improves the *Baseline* from 28.5% to 38.4% in mAP. The proposed *SMF* is evaluated under three settings including with *Splitting* op-

PASCAL VOC → Clipart1k																					
Method	aero	bcyc.	bird	boat	bott.	bus	car	cat	chair	cow	table	dog	horse	bike	pers.	plant	sheep	sofa	train	tv	mAP
Deformable-DETR [81]	24.8	50.5	14.0	22.8	11.5	50.7	28.7	3.0	26.5	32.6	22.1	17.4	19.6	73.1	54.2	20.8	11.5	12.6	55.2	30.3	29.1
DAF [7]	33.5	39.6	24.9	31.4	19.0	61.8	34.5	11.0	29.2	28.5	22.6	20.9	26.5	61.4	51.6	26.7	8.3	23.1	59.7	39.5	32.7
SWDA [51]	38.6	**53.0**	29.4	**39.5**	25.2	**64.8**	36.9	21.4	37.9	39.5	30.7	28.7	31.4	**73.7**	63.4	33.5	15.8	29.2	61.3	41.2	39.8
SCL [55]	32.3	46.8	31.9	36.0	36.8	43.6	40.9	24.4	35.1	37.8	18.1	**34.9**	32.6	67.3	64.5	**43.2**	14.5	30.4	53.5	43.6	38.4
GAP [65]	28.9	42.4	32.4	36.8	36.5	40.8	39.1	23.2	34.6	39.1	16.6	33.1	**36.4**	65.2	66.0	40.1	14.3	30.6	56.4	39.5	37.6
SFA [59]	35.2	47.6	**33.5**	38.3	**39.6**	40.4	38.5	**27.2**	37.6	43.1	23.9	31.6	32.5	72.5	**66.8**	43.0	**18.5**	29.0	53.0	44.9	39.8
DA-DETR	**43.1**	47.7	31.5	33.7	21.4	62.8	**42.6**	14.8	**39.5**	**44.2**	**35.9**	27.5	31.8	72.6	65.6	42.2	17.3	**31.1**	**71.3**	**50.1**	**41.3**

Table 5. Experimental results (%) of the scenario Real-world images to Clipart-style images: PASCAL VOC → Clipart1k.

PASCAL VOC → Watercolor2k							
Method	bike	bird	car	cat	dog	person	mAP
Deformable-DETR [81]	43.3	39.9	21.0	**50.3**	13.7	49.1	36.2
DAF [7]	58.0	41.7	30.2	32.7	34.5	66.9	44.0
SWDA [51]	**58.7**	**53.7**	25.3	40.2	32.8	70.2	46.8
UaDAN [24]	57.2	47.8	31.0	37.8	34.9	70.3	48.2
DA-DETR	58.6	**53.7**	**31.9**	46.2	**40.2**	**73.0**	**50.6**
PASCAL VOC → Comic2k							
Method	bike	bird	car	cat	dog	person	mAP
Deformable-DETR [81]	22.3	13.6	19.6	16.6	18.9	33.1	20.3
DAF [7]	27.8	17.5	28.7	24.5	20.8	45.5	27.5
SWDA [51]	36.6	12.8	**29.5**	16.5	**33.2**	61.7	31.7
UaDAN [24]	37.3	17.3	25.3	**28.5**	29.0	61.9	33.2
DA-DETR	**44.2**	**18.1**	25.0	27.7	33.0	**62.4**	**35.1**

Table 6. Experimental results (%) of the scenarios Real-world images to Watercolor-style images: PASCAL VOC → Watercolor2k and Real-world images to Comic-style images: PASCAL VOC → Comic2k.

eration, with *Shuffling* operation and with both. It can be observed that *SMF* under all three settings outperforms the *Direct-align* consistently, while the SMF with both *Splitting* and *Shuffling* performs the best. In addition, including *SAF* alone over the *Direct-align* improves mAP by 3.3%. The incorporation of *SMF* and *SAF* achieves the best mAP at 43.5%, demonstrating that *SMF* and *SAF* are complementary to each other.

5.4. Comparisons with the State-of-the-Art

We evaluate DA-DETR under four domain-shift scenarios: 1) Normal weather to Foggy weather; 2) Synthetic scene to Real scene; 3) Cross-camera Adaptations and 4) Real-world images to Artistic images. In each domain-shift scenario, we compare DA-DETR with a number of state-of-the-art unsupervised domain adaptive methods.

Normal Weather to Foggy Weather: We first study the adaptation from normal weather to foggy weather on the task Cityscapes → Foggy cityscapes. As Table 2 shows, DA-DETR outperforms the baseline deformable DETR [81] greatly. It also outperforms the state-of-the-art [59] by 2.2% in mAP. For certain categories such as 'train' that can not be well detected by existing methods, DA-DETR achieves the best AP of 37.5. Such experimental results verify that the proposed CTBlender helps to identify both the semantic information and the localization information effectively.

Synthetic Scene to Real Scene: Table 3 shows experiments of adaptation from synthetic to real scenes on the task SIM 10k → Cityscapes. We can observe that DA-DETR achieves the best accuracy with a mAP 54.7%, showing that DA-DETR is powerful when there is only one object category 'car' in cross-domain detection task.

Cross-camera Adaptation: Table 4 shows experiments of cross-camera adaptation over the task KITTI → Cityscapes. We can observe that DA-DETR outperforms the state-of-the-art and improves the baseline model [81] from 39.5% to 48.9% in mAP. These experiments further show that the proposed DA-DETR can well generalize to different domain adaptation tasks.

Real-world Images to Artistic Images: We evaluate the adaptation from real-world images to clipart-style images on the task PASCAL VOC → Clipart1k. Table 5 shows experimental results, where DA-DETR achieves the best mAP of 41.3%. In addition, DA-DETR improves the baseline by large margins for certain categories that are not well detected by the baseline model [81] such as bird and sofa. This experiment shows that DA-DETR can handle domain adaptation with multiple categories effectively.

To demonstrate the generalization capability of DA-DETR, we also evaluate it over the tasks PASCAL VOC → Watercolor2k and PASCAL VOC → Comic2k, respectively. As Table 6 shows, DA-DETR outperforms the baseline [81] by large margins, and it also outperforms all state-of-the-art methods over two tasks consistently.

5.5. Discussion

Effectiveness of Split Fusion in CTBlender. As described in Section 4, the split operation in SMF splits input feature into K groups which helps to encode different semantic information into the fused feature. We examine the effectiveness of the split operation over domain adaptation task Cityscapes → Foggy cityscapes by visualizing the weight generated by each group. We sampled 4 groups from the total 32 groups for each image and highlighted the produced weight over the sample images as shown in Fig. 4. We can observe that SMF captures different foreground regions over different groups, demonstrating that the split operation

Figure 4. Visualization of generated weight in SMF for each group. We take two sample images from validation set of Cityscapes, which are shown in the first column. For each image, we sample 4 groups from the total 32 groups and highlight the generated attention over each sample image as shown in columns 2 to 5, respectively. We can observe that the generated weight in different groups detect different foreground regions effectively. k denotes the k^{th} group defined in Section 4.

Cityscapes → Foggy Cityscapes		
Method	Aligned Features	mAP
Deformable-DETR [81]	N.A.	28.5
Direct-Align [19]	CNN features	38.4
Direct-Align [19]	Transformer features	38.9
Direct-Align [19]	CNN features and Transformer features	40.2
Addition [25]		42.1
Multiplication [61]		41.9
Convolution [72]		41.8
AFF [11]		42.4
LS-DeconvNet [9]		42.6
CTBlender(ours)		**43.5**

Table 7. Comparing the proposed CTBlender with conventional fusion mechanisms in cross-domain alignment (on the domain adaptive object detection task Cityscapes → Foggy Cityscapes).

helps to learn different semantic features effectively.

Analysis of CNN Features and Transformer Features. We study how CNN features and Transformer features affect unsupervised domain adaptation by examining the adaptation performance of the direct alignment of CNN features f, the direct alignment of Transformer features p and both, respectively. As shown in Rows 1-4 of Table 7, aligning CNN features and Transformer features simultaneously brings clear further performance improvement over either CNN feature alignment or Transformer feature alignment. This shows that either the localization information in CNN features or the semantic information in Transformer features can facilitate domain adaptation in some degree while these two types of information are complementary for cross-domain alignment.

Comparison with Conventional Fusion Mechanisms. We study how different feature fusion strategies affect the domain adaptation performance by comparing our CTBlender with existing feature fusion strategies [9, 11, 25, 61, 72], *e.g.*, fusing CNN features and Transformer features via 1) addition [25], multiplication [61] and convolution [72], and 2) attention-based fusion [9, 11]. As shown in the bottom part of Table 7, all fusion strategies improve the *Direct-Align* baseline clearly, demonstrating the effectiveness of aligning the fused features in UDA. In addition, we can observe that our CTBlender performs the best clearly, largely attributed to its split-merge fusion and scale aggregation fusion designs that fuses CNN features and Transformer features ingeniously. Specifically, the split-merge fusion in CTBlender splits the CNN/Transformer feature which enables to fuse them along spatial and channel dimensions respectively, leading to comprehensive information fusion along both feature dimensions. Besides, the scale aggregation fusion in CTBlender aggregates rich information across multiple image scales, leading to effective cross-domain feature alignment for different scales that facilitates cross-domain detection against large scale variance.

6. Conclusion

This paper presents DA-DETR, an unsupervised domain adaptive detection transformer that introduces information fusion into the DETR framework for effective knowledge transfer from a labeled source domain to an unlabeled target domain. We design a novel CNN-Transformer Blender that fuses the CNN features and Transformer features ingeniously for effective feature alignment and domain adaptation across domains. Extensive experiments over multiple domain adaptation scenarios show that DA-DETR achieves superior performance in unsupervised domain adaptive object detection. Moving forwards, we plan to continue to investigate innovative cross-domain alignment strategies for better domain adaptive transformer detection.

Acknowledgement. This study is supported under the RIE2020 Industry Alignment Fund – Industry Collaboration Projects (IAF-ICP) Funding Initiative, as well as cash and in-kind contribution from the industry partner(s).

References

[1] Vinicius F Arruda, Thiago M Paixão, Rodrigo F Berriel, Alberto F De Souza, Claudine Badue, Nicu Sebe, and Thiago Oliveira-Santos. Cross-domain car detection using unsupervised image-to-image translation: From day to night. In *2019 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE, 2019. 2

[2] Léon Bottou. Large-scale machine learning with stochastic gradient descent. In *Proceedings of COMPSTAT'2010*, pages 177–186. Springer, 2010. 6

[3] Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*, 2020. 2

[4] Yue Cao, Jiarui Xu, Stephen Lin, Fangyun Wei, and Han Hu. Gcnet: Non-local networks meet squeeze-excitation networks and beyond. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, pages 0–0, 2019. 2

[5] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European Conference on Computer Vision*, pages 213–229. Springer, 2020. 1, 2, 3

[6] Chaoqi Chen, Zebiao Zheng, Xinghao Ding, Yue Huang, and Qi Dou. Harmonizing transferability and discriminability for adapting object detectors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8869–8878, 2020. 2

[7] Yuhua Chen, Wen Li, Christos Sakaridis, Dengxin Dai, and Luc Van Gool. Domain adaptive faster r-cnn for object detection in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3339–3348, 2018. 2, 5, 6, 7

[8] Bowen Cheng, Alex Schwing, and Alexander Kirillov. Per-pixel classification is not all you need for semantic segmentation. In *Thirty-Fifth Conference on Neural Information Processing Systems*, 2021. 2

[9] Yanhua Cheng, Rui Cai, Zhiwei Li, Xin Zhao, and Kaiqi Huang. Locality-sensitive deconvolution networks with gated fusion for rgb-d indoor semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3029–3037, 2017. 1, 2, 8

[10] 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 *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3213–3223, 2016. 5, 6

[11] Yimian Dai, Fabian Gieseke, Stefan Oehmcke, Yiquan Wu, and Kobus Barnard. Attentional feature fusion. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 3560–3569, 2021. 1, 2, 8

[12] Zhigang Dai, Bolun Cai, Yugeng Lin, and Junying Chen. Up-detr: Unsupervised pre-training for object detection with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1601–1610, 2021. 2

[13] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 6

[14] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. 2

[15] Bin Dong, Fangao Zeng, Tiancai Wang, Xiangyu Zhang, and Yichen Wei. Solq: Segmenting objects by learning queries. *arXiv preprint arXiv:2106.02351*, 2021. 2

[16] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 2

[17] Stéphane d'Ascoli, Hugo Touvron, Matthew L Leavitt, Ari S Morcos, Giulio Biroli, and Levent Sagun. Convit: Improving vision transformers with soft convolutional inductive biases. In *International Conference on Machine Learning*, pages 2286–2296. PMLR, 2021. 1

[18] Mark Everingham, SM Ali Eslami, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes challenge: A retrospective. *International journal of computer vision*, 111(1):98–136, 2015. 6

[19] Yaroslav Ganin and Victor Lempitsky. Unsupervised domain adaptation by backpropagation. In *International conference on machine learning*, pages 1180–1189. PMLR, 2015. 1, 2, 5, 8

[20] Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun. Vision meets robotics: The kitti dataset. *The International Journal of Robotics Research*, 32(11):1231–1237, 2013. 6

[21] Ross Girshick. Fast r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 1440–1448, 2015. 1

[22] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 580–587, 2014. 1

[23] Dayan Guan, Jiaxing Huang, Shijian Lu, and Aoran Xiao. Scale variance minimization for unsupervised domain adaptation in image segmentation. *Pattern Recognition*, 112:107764, 2021. 2

[24] Dayan Guan, Jiaxing Huang, Aoran Xiao, Shijian Lu, and Yanpeng Cao. Uncertainty-aware unsupervised domain adaptation in object detection. *IEEE Transactions on Multimedia*, 2021. 6, 7

[25] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 2, 3, 6, 8

[26] Zhenwei He and Lei Zhang. Multi-adversarial faster-rcnn for unrestricted object detection. In *Proceedings of the*

IEEE/CVF International Conference on Computer Vision, pages 6668–6677, 2019. 2, 5

[27] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7132–7141, 2018. 2

[28] Jiaxing Huang, Dayan Guan, Aoran Xiao, and Shijian Lu. Cross-view regularization for domain adaptive panoptic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10133–10144, 2021. 2

[29] Jiaxing Huang, Dayan Guan, Aoran Xiao, and Shijian Lu. Fsdr: Frequency space domain randomization for domain generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6891–6902, 2021. 2

[30] Jiaxing Huang, Dayan Guan, Aoran Xiao, and Shijian Lu. Model adaptation: Historical contrastive learning for unsupervised domain adaptation without source data. *Advances in Neural Information Processing Systems*, 34:3635–3649, 2021. 2

[31] Jiaxing Huang, Dayan Guan, Aoran Xiao, Shijian Lu, and Ling Shao. Category contrast for unsupervised domain adaptation in visual tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1203–1214, 2022. 2

[32] Jiaxing Huang, Shijian Lu, Dayan Guan, and Xiaobing Zhang. Contextual-relation consistent domain adaptation for semantic segmentation. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XV*, pages 705–722. Springer, 2020. 2

[33] Naoto Inoue, Ryosuke Furuta, Toshihiko Yamasaki, and Kiyoharu Aizawa. Cross-domain weakly-supervised object detection through progressive domain adaptation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5001–5009, 2018. 2, 5, 6

[34] 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? *arXiv preprint arXiv:1610.01983*, 2016. 6

[35] Taekyung Kim, Minki Jeong, Seunghyeon Kim, Seokeon Choi, and Changick Kim. Diversify and match: A domain adaptive representation learning paradigm for object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 12456–12465, 2019. 2, 5

[36] Congcong Li, Dawei Du, Libo Zhang, Longyin Wen, Tiejian Luo, Yanjun Wu, and Pengfei Zhu. Spatial attention pyramid network for unsupervised domain adaptation. In *European Conference on Computer Vision*, pages 481–497. Springer, 2020. 2, 6

[37] Xiang Li, Xiaolin Hu, and Jian Yang. Spatial group-wise enhance: Improving semantic feature learning in convolutional networks. *arXiv preprint arXiv:1905.09646*, 2019. 2

[38] Che-Tsung Lin. Cross domain adaptation for on-road object detection using multimodal structure-consistent image-to-image translation. In *2019 IEEE International Conference on Image Processing (ICIP)*, pages 3029–3030. IEEE, 2019. 2

[39] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2117–2125, 2017. 2

[40] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019. 2

[41] Zhipeng Luo, Zhongang Cai, Changqing Zhou, Gongjie Zhang, Haiyu Zhao, Shuai Yi, Shijian Lu, Hongsheng Li, Shanghang Zhang, and Ziwei Liu. Unsupervised domain adaptive 3d detection with multi-level consistency. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8866–8875, 2021. 2

[42] Zhipeng Luo, Xiaobing Zhang, Shijian Lu, and Shuai Yi. Domain consistency regularization for unsupervised multi-source domain adaptive classification. *Pattern Recognition*, 132:108955, 2022. 2

[43] Ningning Ma, Xiangyu Zhang, Hai-Tao Zheng, and Jian Sun. Shufflenet v2: Practical guidelines for efficient cnn architecture design. In *Proceedings of the European conference on computer vision (ECCV)*, pages 116–131, 2018. 5

[44] Pedro O Pinheiro. Unsupervised domain adaptation with similarity learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8004–8013, 2018. 2

[45] Wang Qilong, Wu Banggu, Zhu Pengfei, Li Peihua, Zuo Wangmeng, and Hu Qinghua. Eca-net: Efficient channel attention for deep convolutional neural networks. 2020. 2

[46] Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. Improving language understanding by generative pre-training. 2018. 2

[47] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019. 2

[48] Maithra Raghu, Thomas Unterthiner, Simon Kornblith, Chiyuan Zhang, and Alexey Dosovitskiy. Do vision transformers see like convolutional neural networks? *Advances in Neural Information Processing Systems*, 34:12116–12128, 2021. 1

[49] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *arXiv preprint arXiv:1506.01497*, 2015. 1

[50] S Hamid Rezatofighi, Vijay Kumar BG, Anton Milan, Ehsan Abbasnejad, Anthony Dick, and Ian Reid. Deepsetnet: Predicting sets with deep neural networks. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 5257–5266. IEEE, 2017. 2

[51] Kuniaki Saito, Yoshitaka Ushiku, Tatsuya Harada, and Kate Saenko. Strong-weak distribution alignment for adaptive object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6956–6965, 2019. 2, 5, 6, 7

[52] Kuniaki Saito, Kohei Watanabe, Yoshitaka Ushiku, and Tatsuya Harada. Maximum classifier discrepancy for unsupervised domain adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3723–3732, 2018. 2

[53] 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. 6

[54] Yuhu Shan, Wen Feng Lu, and Chee Meng Chew. Pixel and feature level based domain adaptation for object detection in autonomous driving. *Neurocomputing*, 367:31–38, 2019. 2

[55] Zhiqiang Shen, Harsh Maheshwari, Weichen Yao, and Marios Savvides. Scl: Towards accurate domain adaptive object detection via gradient detach based stacked complementary losses. *arXiv preprint arXiv:1911.02559*, 2019. 2, 6, 7

[56] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. *arXiv preprint arXiv:1706.03762*, 2017. 2

[57] Vibashan VS, Vikram Gupta, Poojan Oza, Vishwanath A Sindagi, and Vishal M Patel. Mega-cda: Memory guided attention for category-aware unsupervised domain adaptive object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4516–4526, 2021. 2

[58] Tuan-Hung Vu, Himalaya Jain, Maxime Bucher, Matthieu Cord, and Patrick Pérez. Advent: Adversarial entropy minimization for domain adaptation in semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2517–2526, 2019. 2

[59] Wen Wang, Yang Cao, Jing Zhang, Fengxiang He, Zheng-Jun Zha, Yonggang Wen, and Dacheng Tao. Exploring sequence feature alignment for domain adaptive detection transformers. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 1730–1738, 2021. 2, 6, 7

[60] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7794–7803, 2018. 2

[61] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. Cbam: Convolutional block attention module. In *Proceedings of the European conference on computer vision (ECCV)*, pages 3–19, 2018. 2, 8

[62] Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M Alvarez, and Ping Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. *arXiv preprint arXiv:2105.15203*, 2021. 2

[63] Yun Xing, Dayan Guan, Jiaxing Huang, and Shijian Lu. Domain adaptive video segmentation via temporal pseudo supervision. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXX*, pages 621–639. Springer, 2022. 2

[64] Chang-Dong Xu, Xing-Ran Zhao, Xin Jin, and Xiu-Shen Wei. Exploring categorical regularization for domain adaptive object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11724–11733, 2020. 2, 5, 6

[65] Minghao Xu, Hang Wang, Bingbing Ni, Qi Tian, and Wenjun Zhang. Cross-domain detection via graph-induced prototype alignment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12355–12364, 2020. 6, 7

[66] Yanchao Yang and Stefano Soatto. Fda: Fourier domain adaptation for semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4085–4095, 2020. 2

[67] Fuxun Yu, Di Wang, Yinpeng Chen, Nikolaos Karianakis, Pei Yu, Dimitrios Lymberopoulos, and Xiang Chen. Unsupervised domain adaptation for object detection via cross-domain semi-supervised learning. *arXiv preprint arXiv:1911.07158*, 2019. 2

[68] Jinze Yu, Jiaming Liu, Xiaobao Wei, Haoyi Zhou, Yohei Nakata, Denis Gudovskiy, Tomoyuki Okuno, Jianxin Li, Kurt Keutzer, and Shanghang Zhang. Mttrans: Cross-domain object detection with mean teacher transformer. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part IX*, pages 629–645. Springer, 2022. 2, 6

[69] Li Yuan, Yunpeng Chen, Tao Wang, Weihao Yu, Yujun Shi, Zi-Hang Jiang, Francis EH Tay, Jiashi Feng, and Shuicheng Yan. Tokens-to-token vit: Training vision transformers from scratch on imagenet. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 558–567, 2021. 1

[70] Gongjie Zhang, Zhipeng Luo, Kaiwen Cui, Shijian Lu, and Eric P Xing. Meta-detr: Image-level few-shot detection with inter-class correlation exploitation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022. 2

[71] Gongjie Zhang, Zhipeng Luo, Yingchen Yu, Zichen Tian, Jingyi Zhang, and Shijian Lu. Towards efficient use of multi-scale features in transformer-based object detectors. *arXiv preprint arXiv:2208.11356*, 2022. 2

[72] Hang Zhang, Chongruo Wu, Zhongyue Zhang, Yi Zhu, Haibin Lin, Zhi Zhang, Yue Sun, Tong He, Jonas Mueller, R Manmatha, et al. Resnest: Split-attention networks. *arXiv preprint arXiv:2004.08955*, 2020. 4, 8

[73] Jingyi Zhang, Jiaxing Huang, Zichen Tian, and Shijian Lu. Spectral unsupervised domain adaptation for visual recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9829–9840, 2022. 2

[74] Pan Zhang, Bo Zhang, Ting Zhang, Dong Chen, Yong Wang, and Fang Wen. Prototypical pseudo label denoising and target structure learning for domain adaptive semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12414–12424, 2021. 2

[75] Xiangyu Zhang, Xinyu Zhou, Mengxiao Lin, and Jian Sun. Shufflenet: An extremely efficient convolutional neural network for mobile devices. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6848–6856, 2018. 5

[76] Yixin Zhang, Zilei Wang, and Yushi Mao. Rpn prototype alignment for domain adaptive object detector. In *Proceed-*

ings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 12425–12434, 2021. 2

[77] Hengshuang Zhao, Li Jiang, Jiaya Jia, Philip HS Torr, and Vladlen Koltun. Point transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 16259–16268, 2021. 2

[78] Sixiao Zheng, Jiachen Lu, Hengshuang Zhao, Xiatian Zhu, Zekun Luo, Yabiao Wang, Yanwei Fu, Jianfeng Feng, Tao Xiang, Philip HS Torr, et al. Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6881–6890, 2021. 2

[79] 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. 2, 6

[80] Xinge Zhu, Jiangmiao Pang, Ceyuan Yang, Jianping Shi, and Dahua Lin. Adapting object detectors via selective cross-domain alignment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 687–696, 2019. 2

[81] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable detr: Deformable transformers for end-to-end object detection. *arXiv preprint arXiv:2010.04159*, 2020. 1, 2, 3, 6, 7, 8

[82] Yang Zou, Zhiding Yu, BVK Kumar, and Jinsong Wang. Unsupervised domain adaptation for semantic segmentation via class-balanced self-training. In *Proceedings of the European conference on computer vision (ECCV)*, pages 289–305, 2018. 2

[83] Yang Zou, Zhiding Yu, Xiaofeng Liu, BVK Kumar, and Jinsong Wang. Confidence regularized self-training. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 5982–5991, 2019. 2