DA-DETR: Domain Adaptive Detection Transformer by Hybrid Attention

Jingyi Zhang†    Jiaxing Huang†    Zhipeng Luo    Gongjie Zhang    Shijian Lu∗
Nanyang Technological University, Singapore
jingyi.zhang@ntu.edu.sg jiaxing.huang@ntu.edu.sg zhipeng001@e.ntu.edu.sg

gongjiezhang@ntu.edu.sg shijian.lu@ntu.edu.sg

Abstract

The prevalent approach in domain adaptive object detection adopts a two-stage architecture (Faster R-CNN) that involves a number of hyper-parameters and hand-crafted designs such as anchors, region pooling, non-maximum suppression, etc. Such architecture makes it very complicated while adopting certain existing domain adaptation methods with different ways of feature alignment. In this work, we adopt a one-stage detector and design DA-DETR, a simple yet effective domain adaptive object detection network that performs inter-domain alignment with a single discriminator. DA-DETR introduces a hybrid attention module that explicitly pinpoints the hard-aligned features for simple yet effective alignment across domains. It greatly simplifies traditional domain adaptation pipelines by eliminating sophisticated routines that involve multiple adversarial learning frameworks with different types of features. Despite its simplicity, extensive experiments show that DA-DETR demonstrates superior accuracy as compared with highly-optimized state-of-the-art approaches.

1. Introduction

Object detection has been a longstanding challenge in computer vision, which aims to assign a bounding box and a class label to every object inside an image. Deep learning based methods [15, 14, 38, 7, 31, 27, 11, 33, 35, 54, 2, 5, 58] have achieved great success at the price of large quantities of annotated training data [32, 12, 9] that are prohibitively expensive and time-consuming to collect [32, 9, 53, 55]. One way of circumventing this constraint is to leverage the off-the-shelf labeled data from a different but related “source domain” [9, 42, 24] in network training. However, such trained models often experience clear performance drops while applied to a “target domain” due to distribution gaps between the two domains [40, 46, 41, 8, 39, 22, 20, 21].

Unsupervised domain adaptation (UDA) has been investigated to address the domain gap issue. Most existing works [8, 39, 17, 50, 59, 3, 51] are based on a complex two-stage architecture (i.e., Faster R-CNN [38]) that comes with a number of heuristic and hand-crafted designs such as anchor generation, region-of-interest pooling, non-maximum suppression, etc. To cater to such architecture, they involve multiple objectives with a number of adversarial discriminators for aligning different types of features, e.g., image-level features [8, 50] (i.e., output of backbone), instance-
level features [8, 50] (i.e., output of region pooling), hierarchical features [17, 39] (i.e., output of multiple network layers) as illustrated in Fig. 1. Although quite impressive progress has been achieved, they complicate the network design, cannot fully harvest synergetic relations among different network components, and often lead to uncoordinated network training and sub-optimal detection models.

We adopt a one-stage detector and design DA-DETR, a simple yet effective Domain Adaptive DEtection TRansformer that performs inter-domain alignment with a single discriminator. DA-DETR introduces a hybrid attention module (HAM) that explicitly pinpoints hard-aligned features for simple yet effective alignment across domains as illustrated in Fig. 1. HAM consists of two sequential modules, including a coordinate attention module (CAM) that embeds positional information into channel attention to hunt for hard-aligned target features and a level attention module (LAM) that aggregates attention features across multiple scales at deformation levels. Different from traditional stand-alone spatial and channel attention [18, 48], CAM splits features from backbone into two parts and fuses them with latent features from transformer encoder to capture rich contextual and positional information. The produced features are then concatenated and shuffled to facilitate information flow across channels. With this design, HAM can explicitly pinpoint hard-aligned features and enables straight inter-domain alignment with a single discriminator. Extensive experiments show that DA-DETR greatly simplifies domain adaptive detection pipelines and produces superior detection over multiple benchmarks.

The contributions of this work can be summarized in three aspects. First, we propose DA-DETR, a simple yet effective domain adaptive detection network that achieves superior feature alignment with a single discriminator. Second, we design a hybrid attention module that automatically pinpoints hard-aligned target features and aligns them across domains effectively. Third, extensive experiments over multiple domain adaptation benchmarks show that DA-DETR outperforms the state-of-the-art consistently.

2. Related Works

Object Detection. The advance of CNNs [26, 45, 16, 19] has improved object detection greatly in recent years. State-of-the-art approaches can be broadly grouped into two categories, namely, two-stage detectors and one-stage detectors. Specifically, two-stage detectors [15, 14, 38] handle two sequential tasks on proposal generation and proposal prediction, whereas one-stage detectors [33, 35, 36, 37, 2, 11, 27] removes the proposal generation stage and directly locate objects in images. Recently, transformer-based detectors have been investigated which eliminate hand-crafted designs such as non-maximum suppression and anchor generation and achieve fully end-to-end detection. For example, DETR [5] first adopts the transformer architecture [47] for the object detection task. Deformable DETR [58] extends DETR with a deformable attention module that reduces the training time significantly. However, how to leverage the transformer architecture in domain adaptive object detection has not been explored to the best of our knowledge.

Unsupervised Domain Adaptation (UDA). The target of UDA in object detection 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 [8, 39, 56, 28, 44, 50, 57, 6] adopt Faster R-CNN and strive to align different type of features (e.g., features from backbone, instance-level features from ROI pooling, etc.) via adversarial learning [13]. For example, DA Faster R-CNN [8], the first work tackling UDA in object detection, employs consistency regularization to harvest the synergetic relations among multiple feature alignment pipelines. In addition, a few studies [1, 25, 43, 52, 1, 30] translate data to mitigate domain shift at the input level, e.g., [52] builds an intermediate domain that is more similar to the target domain, and adopts self-training for adaptation. Different from the sophisticated routines of previous approaches, our proposed DA-DETR can realize a comprehensive feature alignment across source and target domains with a single discriminator and achieve a superior performance.

Attention Mechanism. Attention mechanism has been widely studied in various computer vision tasks for guiding models to focus on informative image regions. Most existing attention mechanisms can be broadly classified into spatial attention, channel attention and the combination of the two. Specifically, spatial attention [48] aims to capture spatial dependencies in an image, while channel attention [18, 34] studies the relation of convolutional features of different channels. Recently, a few studies such as CBAM [49], GCNet [4] and SGE [29] strive to integrate spatial attention and channel attention for better focus on informative regions. Different from previous works, we design a hybrid attention module that fuses multi-source attention optimally for the domain adaptive object detection task.

3. Method

This section presents the proposed domain adaptive detection transformer (DA-DETR) that introduces a hybrid attention module for optimal feature alignment with a single discriminator as illustrated in Fig. 2. It consists of five subsections that focus on Task definition, Framework Overview, Coordinate Attention Module (CAM) that embeds the positional information into channel attention to hunt for hard-aligned target features, Level Attention Module (LAM) that aggregates attention features across multiple scales at deformation levels, and Network Training, respectively.
3.1. Task Definition

The work focuses on the problem of unsupervised domain adaptation (UDA) in object detection. It involves a source domain $D_s$ and a target domain $D_t$, where $D_s = \{(x_s^i, y_s^i)\}_{i=1}^{N_s}$ is fully labelled, 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 unlabelled target-domain data $x_t^i$. The baseline model is trained with the labelled source data (i.e., $D_s$) only:

$$L_{det} = l(T(G(x_s)), y_s),$$  

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

3.2. Framework Overview

As shown on the top of 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 hybrid attention module (HAM) $H$. We adopt the deformable DETR 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. HAM consists of two sub-modules including a coordinate attention module (CAM) and a level attention module (LAM). Taking the features from $G$ and the positional information from the encoder $E$ of $T$ as inputs, HAM encodes the positional information to identify hard-aligned target features among the features generated by $G$.

Given a source image $x_s \in D_s$ and a target image $x_t \in D_t$, the backbone $G$ will first produce feature maps $f_s$ and $f_t$, respectively. The backbone features are then fed to the transformer encoder $E$ to obtain latent features $p_s$ and $p_t$. HAM then takes the backbone features $f_s$ and $f_t$ and the latent features $p_s$ and $p_t$ to pinpoint hard-aligned features $V_s^a$ and $V_t^a$ among $f_s$ and $f_t$. Finally, the identified $V_s^a$ and $V_t^a$ are fed to the discriminator $C_d$ to generate an adversarial loss $L_{adv}$ for inter-domain feature alignment. For the source data flow, the latent feature $p_s$ is also fed to the transformer decoder $D$ to predict a set of bounding boxes and pre-defined semantic categories, which will be used to calculate a detection loss $L_{det}$ under the supervision of the ground-truth label $y_s \in D_s$. The overall network is optimized by the adversarial loss $L_{adv}$ and the detection loss $L_{det}$.

3.3. Coordinate Attention Module

The coordinate attention module (CAM) within the hybrid attention module (HAM) fuses spatial attention and channel attention (computed with the latent features $p$ from the transformer encoder) for optimally re-weighting features $f$. As the transformer encoder consists of a stack of self-attention components, the latent features $p$ capture rich positional information of the input image and can reflect the dependencies among all image pixels well. Therefore, $p$ can be exploited to guide to pinpoint hard-aligned features effectively. In our network, we feed the backbone features $f = \{f^l\}_{l=1}^{L}$ and latent features $p = \{p^l\}_{l=1}^{L}$ to the CAM ($l$ denotes the $l$-th deformation level and $L = 4$ in deformable DETR).

**Split Attention.** As CAM operation for each deformation level is the same, we take the first level $l = 1$ as an example to illustrate how we perform coordinate attention. Given a latent feature $p^1 \in \mathbb{R}^{C \times H \times W}$ and a backbone 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 along channels, i.e., $\{p^1_k\}_{k=1}^{K} \in \mathbb{R}^{(C/K) \times H \times W}$, where each group captures different semantic information of the input image. To perform coordinate attention, we further split each group of feature $p^1_k$ into two parts along channels equally, i.e., $p^1_{k1} \in \mathbb{R}^{(C/2K) \times H \times W}$ and $p^1_{k2} \in \mathbb{R}^{(C/2K) \times H \times W}$. The split features will be used to generate spatial attention and channel attention.

For the spatial attention generation, the feature $p^1_{k1} \in \mathbb{R}^{(C/2K) \times H \times W}$ is firstly fed into a Group Normalization (GN) layer and then re-weighted by a learnable weight map $w_s \in \mathbb{R}^{(C/2K) \times H \times W}$ and a learnable bias map $b_s \in \mathbb{R}^{(C/2K) \times H \times W}$:

$$p^1_{k1} = f_s (w_s \cdot GN(p^1_{k1}) + b_s),$$

where $f_s(\cdot)$ is an activation function to limit the input in $[0, 1]$.

For the channel attention generation, the feature $p^1_{k2} \in \mathbb{R}^{(C/2K) \times H \times W}$ is firstly compacted by the Global Average Pooling (GAP):

$$v^1_{k2} = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} p^1_{k2}(i, j),$$

where $v^1_{k2} \in \mathbb{R}^{(C/2K) \times 1 \times 1}$.

Similar to spatial attention, $v^1_{k2}$ is then re-weighted by a learnable weight vector $w_c \in \mathbb{R}^{(C/2K) \times 1 \times 1}$ and a learnable bias vector $b_c \in \mathbb{R}^{(C/2K) \times 1 \times 1}$ to assign different weights to different elements inside $v^1_{k2}$:

$$\hat{p}^1_{k2} = f_c (w_c \cdot v^1_{k2} + b_c),$$

where $f_c(\cdot)$ is an activation function to limit $\hat{p}^1_{k2}$ in $[0, 1]$.

Similar to the operation for latent feature $p^1$, the backbone feature $f^1$ is also divided into $K$ groups along the channels, i.e., $\{f^1_k\}_{k=1}^{K} \in \mathbb{R}^{(C/K) \times H \times W}$. Each divided
Figure 2. Overview of the proposed DA-DETR: We design a novel hybrid attention module (HAM) for domain adaptive object detection. HAM consists of a coordinate attention module (CAM) and a level attention module (LAM) as illustrated. In CAM, latent features from the transformer encoder of all deformation levels $p_l=1,2,3,4$ are adopted to generate attention to modulate the backbone features $f_l=1,2,3,4$. Take the first deformation level $f_1$, $p_1$ as example. The latent feature $p_1$ is divided into $K$ groups (e.g., $K=2$) and further split into two parts to generate spatial attention by $A_s$ and channel attention by $A_c$. Similarly, the backbone output $f_1$ is divided and split in a same way, and then weighted by the spatial attention and channel attention, respectively. The weighted features (e.g., $\hat{f}_1$ and $\hat{f}_2$) are concatenated and shuffled to generate final weighted features for the first level. In LAM, $f_l=1,2,3,4$ are re-weighted by level attention, where GAP denotes global average pooling, and weight coefficients are generated by $A_l$. The final feature $V$ is obtained by an element-wise addition.

Shuffle Attention. We propose shuffle attention to enable information communication across channels. Specifically, we first re-weight the split feature $f_{k1}/f_{k2}$ by the corresponding re-weighted latent feature $\hat{p}_{k1}/\hat{p}_{k2}$:

$$\hat{f}_k = f_c((f_{k1} \cdot \hat{p}_{k1}) , (f_{k2} \cdot \hat{p}_{k2})),$$

where function $f_c()$ denotes the concatenation of tensors along channels. The spatial attention and attention is thus embedded into $f_k$. As the spatial and channel attention is independent in the combined feature $f_k$, we shuffle $\hat{f}_k$ along channels to enable information flow across channels for better identification of hard-aligned features.

Lastly, we conduct the above operations $K$ times to generate $K$ shuffled attention features for each group, i.e.,

$$\{\hat{f}_1\}_{K}^{K} \in \mathbb{R}^{(C/K) \times H \times W},$$

where similar operations are conducted to get the results of all levels $\hat{f} = \{\hat{f}_1\}_{l=1}^{L}$.

3.4. Level Attention Module

The previous section describes CAM that re-weights $f = \{f_l\}_{l=1}^{L}$ to identify hard-aligned features $\hat{f} = \{\hat{f}_l\}_{l=1}^{L}$ at each level. To explicitly identify hard-aligned features of different scales (e.g., regional and global context features), we design a level attention module (LAM) to aggregate features $\hat{f}$ across scales at deformation levels.
Specifically, we compact each level of feature \( \hat{f} = \{\hat{f}^l\}_{l=1}^{L} \) into a channel-wise vector \( u = \{u^l\}_{l=1}^{L} \in \mathbb{R}^{C \times 1 \times 1} \) via a Global Average Pooling (GAP) layer. The level coefficients \( \alpha_l \) are obtained from channel-wise vectors \( u^l \) by two steps: merge and split. Firstly, the channel-wise vectors are merged together to obtain merged vector \( u_m \) by an element-wise addition:
\[
 u_m = \sum_{l=1}^{L} u^l, \tag{7}
\]
where \( u_m \in \mathbb{R}^{C \times 1 \times 1} \).

Then, a fully connected layer separates \( u_m \) to \( L \) level coefficient vectors \( \alpha^l \in \mathbb{R}^{C \times 1 \times 1} \). Finally, \( V^\alpha \) is obtained by
\[
 V^\alpha = \sum_{l=1}^{L} \hat{f}^l \cdot \alpha^l, \tag{8}
\]
where \( V^\alpha \) is a highly embedded feature which contains spatial-wise and channel-wise information of the image.

### 3.5. Network Training

This subsection presents how our proposed DA-DETR achieves cross-domain alignment of hard-aligned features that are identified by the hybrid attention module \( \mathcal{H} \). The network is trained with two losses, i.e., a supervised object detection loss \( \mathcal{L}_{\text{det}} \) as defined in Eq. 1 and an adversarial alignment loss \( \mathcal{L}_{\text{adv}} \) that is defined as follow:
\[
 \mathcal{L}_{\text{adv}} = \mathbb{E}_{(f,p) \in \mathcal{D}} \log C_d(\mathcal{H}(f,p)) + \mathbb{E}_{(f,p) \in \mathcal{D}} \log (1 - C_d(\mathcal{H}(f,p))), \tag{9}
\]
where \( f = G(x) \) and \( p = E(G(x)) \). \( G \) denotes backbone; \( E \) denotes transformer encoder; \( \mathcal{H} \) denotes hybrid attention module (HAM) 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
\[
 \max_{C_t} \min_{G,T,\mathcal{H}} \mathcal{L}_{\text{det}}(G,T) - \lambda \mathcal{L}_{\text{adv}}(\mathcal{H},C_d), \tag{10}
\]
where \( T \) denotes the transformer in DETR; \( \lambda \) denotes the weight factor of adversarial loss \( \mathcal{L}_{\text{adv}} \), which balances the influences of \( \mathcal{L}_{\text{det}} \) and \( \mathcal{L}_{\text{adv}} \) in training. Noted we adopt a gradient reverse layer (GRL) to enable the gradient of \( \mathcal{L}_{\text{adv}} \) to be reversed before back-propagating to \( \mathcal{H} \) from \( C_d \).

### 4. Experiments

In this section, we presents experiments including experiment setups, implementation details, ablation studies, comparisons with the state-of-the-art and discussion, more details to be described in the ensuring subsections.

| Direct-align | +CAM | +LAM | mAP |
|--------------|------|------|-----|
| w/o Sp.      | ✓    | ✓    | 34.0 |
| w/o Sh.      | ✓    | ✓    | 40.5 |
| w/ Sp.&Sh.   | ✓    | ✓    | 41.4 |

| Domain Adaptation Scenarios | Cityscapes → Foggy Cityscapes | Cityscapes → Foggy Cityscapes | Cityscapes → Foggy Cityscapes |
|----------------------------|--------------------------------|--------------------------------|--------------------------------|
| ✓                          | ✓                              | ✓                              | 43.5                           |

| Direct-align | +CAM | +LAM | mAP |
|--------------|------|------|-----|
| w/o Sp.      | ✓    | ✓    | 50.5 |
| w/o Sh.      | ✓    | ✓    | 52.5 |
| w/ Sp.&Sh.   | ✓    | ✓    | 52.9 |

| Dataset Names | Cityscapes | Foggy Cityscapes | Synthetic scene to Real scene |
|---------------|------------|-----------------|-----------------------------|
| Cityscapes    | ✓          | ✓               | ✓                           |
| Foggy Cityscapes | ✓        | ✓               | ✓                           |
| Synthetic scene | ✓         | ✓               | ✓                           |
| Real scene     | ✓          | ✓               | ✓                           |

### 4.1. Experiment Setups

**Datasets.** The evaluations were performed over three datasets including 1) Cityscapes [9] which was collected for the understanding of street scenes. Its images are captured under normal weather conditions from 50 cities, including 2,975 training images and 500 validation images with pixel-wise instance annotations of 8 categories. We follow prior works [8, 39] to generate bounding boxes from the pixel-wise instance annotations; 2) Foggy Cityscapes [42] which is a synthetic dataset that is derived from Cityscapes by adding simulated fog; and 3) SIM 10k [24] which is a synthetic dataset collected from the computer game Grand Theft Auto V (GTA5). This dataset contains 10,000 annotated images with the category of ‘Car’.

**Domain Adaptation Scenarios.** We evaluate DA-DETR under two widely adopted adaptation scenarios including 1) Normal weather to Foggy weather (Cityscapes → Foggy Cityscapes) which aims to achieve domain adaptation across different weather conditions where Cityscapes is used as the source and Foggy Cityscapes is used as the target. The training images of both datasets are used in training and the adaptation is evaluated over the validation set of Foggy Cityscapes; and 2) Synthetic scene to Real scene (SIM 10k → Cityscapes) where the training images of SIM 10k dataset is used as the source and the training images of Cityscapes is used as the target. Evaluations are performed
| Method       | Backbone | person | rider | car  | truck | bus  | train | mcycle | bicycle | mAP  |
|--------------|----------|--------|-------|------|-------|------|-------|--------|---------|------|
| Faster R-CNN [38] | ResNet-50 | 26.9   | 38.2  | 35.6 | 18.3  | 32.4 | 9.6   | 25.8   | 28.6    | 26.9 |
| DAF [8]      | ResNet-50 | 29.2   | 40.4  | 43.4 | 19.7  | 38.3 | 28.5  | 23.7   | 32.7    | 32.0 |
| SWDA [39]    | ResNet-50 | 31.8   | 44.3  | 48.9 | 21.0  | 43.8 | 28.0  | 28.9   | 35.8    | 35.3 |
| DETR [58]    | ResNet-50 | 43.7   | 38.0  | 57.2 | 15.2  | 34.7 | 14.4  | 26.1   | 42.4    | 34.0 |
| DAF [8]      | ResNet-50 | 49.4   | 49.7  | 62.1 | 23.6  | 43.8 | 21.6  | 31.3   | 43.1    | 40.6 |
| SWDA [39]    | ResNet-50 | 49.0   | 49.0  | 61.4 | 23.9  | 43.1 | 22.9  | 31.0   | 45.2    | 40.7 |
| CRDA [50]    | ResNet-50 | 49.8   | 48.4  | 61.9 | 22.3  | 40.7 | 30.0  | 29.9   | 45.4    | 41.1 |
| CF [56]      | ResNet-50 | 49.6   | 49.7  | 62.6 | 23.3  | 43.4 | 27.4  | 30.2   | 44.8    | 41.4 |
| SAP [28]     | ResNet-50 | 49.3   | 49.9  | 62.5 | 23.0  | 44.1 | 29.4  | 31.3   | 45.8    | 41.9 |
| DA-DETR      | ResNet-50 | 49.9   | 50.0  | 63.1 | 24.0  | 45.8 | 37.5  | 31.6   | 46.3    | 43.5 |

Table 3. Experimental results (%) of the scenario Normal weather to Foggy weather: Cityscapes → Foggy Cityscapes.

Figure 3. Qualitative illustration of domain adaptive detection for Cityscapes→Foggy Cityscapes: DA-DETR can adapt well from normal to foggy weather conditions. The DETR [58] and SAP [28] do not take full advantage of different types of attention, which leads to suboptimal cross-domain alignment and object detection under foggy weather conditions.

4.2. Implementation Details

In all experiments, we adopt deformable DETR [58] with ResNet-50 [16] backbone (pre-trained on ImageNet [10]) as the base detector. For adaptation scenario Normal weather to Foggy weather, we train the network with SGD for 50 epochs. In the first 40 epochs, the learning rate is 0.0002 and then 0.00002 for another 10 epochs. For adaptation scenario Synthetic scene to Real scene, we train our network with SGD for 40 epochs. The learning rate is 0.0001 for first 30 epochs and then 0.000001 for another 10 epochs. The weight factor $\lambda$ in Eq. 10 is set at 0.1 and the number of split groups $K$ in CAM is fixed at 32. We report mAP with an IoU threshold of 0.5 in evaluations.

4.3. Ablation Studies

The proposed hybrid attention module (HAM) consists of a coordinate attention module (CAM) and a level attention module (LAM). We studied the two modules to examine how they contribute to the unsupervised domain adaptive detection. Table 1 shows experimental results over the validation data of Foggy Cityscapes under the scenario “normal weather to foggy weather”. Table 2 shows experimental results over the validation data of Cityscapes under the scenario “synthetic scene to real scene”.

As Table 1 shows, the Baseline model (trained with su-
Figure 4. Visualization of attention generated by our hybrid attention module (HAM). We take four sample images from the validation set of Foggy Cityscapes. The sample images are shown in columns 1 and 3, and the computed attention is highlighted over the sample images as shown in columns 2 and 4. We can observe that the HAM-produced attention detects hard-aligned image regions accurately.

| Method      | Backbone | mAP on Car |
|-------------|----------|------------|
| Faster R-CNN [38] | ResNet-50 | 34.6       |
| DAF [8]      | ResNet-50 | 41.9       |
| SWDA [39]    | ResNet-50 | 44.6       |
| DETR [58]    | ResNet-50 | 50.5       |
| DAF [8]      | ResNet-50 | 51.8       |
| SWDA [39]    | ResNet-50 | 51.5       |
| CRDA [50]    | ResNet-50 | 52.1       |
| CF [56]      | ResNet-50 | 52.5       |
| SAP [28]     | ResNet-50 | 53.1       |
| DA-DETR      | ResNet-50 | 55.3       |

Table 4. Experimental results (%) of the scenario Synthetic scene to Real scene: SIM 10k → Cityscapes.

| Scenarios | K (the number of groups in CAM module) | 1 | 4 | 8 | 16 | 32 | 64 |
|-----------|--------------------------------------|---|---|---|----|----|----|
| Weather   | 41.7 | 42.0 | 41.5 | 43.1 | **43.5** | 41.6 |
| Scene     | 52.9 | 53.4 | 53.9 | 54.1 | **55.3** | 54.7 |

Table 5. The sensitivity of parameter $K$ affects domain adaptation in scenarios Normal weather to Foggy weather: Cityscapes → Foggy Cityscapes and Synthetic scene to Real scene: SIM 10k → Cityscapes. The two scenarios are denoted by ‘weather’ and ‘scene’ respectively. The experiments results (%) with $K$ as evaluated in mAP.

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

Since there is few prior research on transformer-based domain adaptation for object detection, we compare the proposed DA-DETR with several Faster R-CNN based domain adaptation methods [8, 39, 50, 56, 28] that achieved state-of-the-art detection performance as shown in the top parts of Tables 3 and 4. In addition, we also adapt these Faster R-CNN approaches to the transformer-based domain adaptive detection for fair comparisons. The adaptation is performed by keeping the domain adaptation modules [8, 39, 50, 56, 28] unchanged but replacing their post-processing modules (e.g., region proposal network, proposal classification module, etc.) by the encoder-decoder module in DETR. The detection performance of the adapted methods (using the same baselines with the proposed DA-DETR) is shown in the bottom parts of Tables 3 and 4.

In addition, we performed the above comparison experiments over two domain adaptation tasks including Cityscapes → Foggy Cityscapes and SIM 10k → Cityscapes. Tables 3 and 4 show all experimental results. As the two tables show, either Faster R-CNN or DETR does not perform well though DETR performs clearly better as it works at image level capturing rich attention and context information. By including domain adaptation, all methods using either Faster R-CNN and DETR improve by large margins, shows their effectiveness in mitigating cross-domain gaps. In addition, the proposed DA-DETR outperforms all state-of-the-art domain adaptation methods clearly and consistently across two adaptation tasks. The outstanding performance is largely attributed to our designed hybrid attention module which captures more context features at image level and identifies hard-aligned features with better cross-
domain alignment. The qualitative comparison is illustrated in Fig. 3.

### 4.5. Discussion

**Parameter Analysis.** The parameter $K$ is important which determines the number of split groups in our coordinate attention module (CAM). We studied the sensitivity of parameter $K$ by changing it from 1 to 64. Since the number of channels ($C$) should be divisible by the number of groups $K$, we set $K = (1, 4, 8, 16, 32, 64)$ in this paper, where $K = 1$ means no splitting and $K = 64$ means each group of feature contains 4 channels. The experiments are conducted over the tasks Cityscapes → Foggy Cityscapes and SIM 10k → Cityscapes, and Table 5 shows experiment results. It can be seen that the detection performance is quite tolerant to the parameter $K$ and the best performance is obtained when $K = 32$.

**Hybrid vs. Conventional Attention.** The hybrid attention module (HAM) in the proposed DA-DETR plays an important role in achieving effective cross-domain alignment, where the visualization of attention generated by HAM is illustrated in Fig.6. To demonstrate how HAM helps to mitigate cross-domain gaps in an effective and unique way, we compare HAM with most popular attention methods by applying them to domain adaptation, where the experiments include direct alignment [13] (without any attention operation), direct alignment with conventional spatial attention [48], direct alignment with conventional channel attention [18] and direct alignment with conventional spatial and channel attention [49]. The feature alignment is performed with two types of features including latent features as produced by the transformer encoder and normal features as produced by the backbone feature extractor.

Table 6 shows experimental results over the task Cityscapes → Foggy Cityscapes. It can be seen that the baseline DETR without any alignment does not perform well due to the domain shift. The simple Direct-Align without involving any attention mechanism outperforms the baseline greatly within both feature spaces. In addition, further involving conventional spatial attention, channel attention or both outperforms the Direct-Align consistently, demonstrating their effectiveness in cross-domain alignment. It can also be seen that all alignment methods perform better in the latent feature space.

The proposed HAM performs clearly better than all compared attention mechanisms in cross-domain alignment. The outstanding performance is attributed to three major factors. First, HAM exploits spatial and channel attention with latent features from the transformer encoder which capture rich image-level and region-level context information. Second, HAM optimally coordinates and fuses spatial and channel attention by channel splitting and shuffling which enables information communication and flow between spatial and channel attentions and accordingly benefits from their synergic relations. Third, HAM naturally captures attention information across multiple scales by leveraging the deformable DETR architecture. The finally produced hybrid attention is fed to re-weight the backbone CNN features to explicitly pinpoint the hard-aligned features for effective cross-domain alignment.

### 5. Conclusion

This paper presents DA-DETR, an unsupervised domain adaptive detection transformer network that achieves simple yet effective feature alignment between source and target domains by using a single discriminator. The proposed DA-DETR has two unique characteristics. First, it simplifies the domain adaptation pipeline by eliminating sophisticated routines that involve multiple discriminators and different types of feature alignment. Second, it pinpoints the hard-aligned features explicitly by a simple hybrid attention module (HAM) that aligns features across domains effectively. Extensive experiments over multiple domain adaptation scenarios show that DA-DETR achieves superior performance in unsupervised domain adaptive object detection. Moving forwards, we will continue to investigate innovative cross-domain alignment strategies for better domain adaptive object detection.
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] Alexey Bochkovskiy, Chien-Yao Wang, and Hong-Yuan Mark Liao. Yolov4: Optimal speed and accuracy of object detection. arXiv preprint arXiv:2004.10934, 2020.

[3] Qi Cai, Yingwei Pan, Chong-Wah Ngo, Xinmei Tian, Lingyu Duan, and Ting Yao. Exploring object relation in mean teacher for cross-domain detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 11457–11466, 2019.

[4] Yue Cao, Jiariu Xu, Stephen Lin, Fangyuan Wei, and Han Hu. Genet: 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.

[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.

[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.

[7] Tao Chen, Shijian Lu, and Jiayuan Fan. S-cnn: Subcategory-aware convolutional networks for object detection. IEEE transactions on pattern analysis and machine intelligence, 40(10):2522–2528, 2017.

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

[9] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 3213–3223, 2016.

[10] 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.

[11] Kaiwen Duan, Song Bai, Lingxi Xie, Honggang Qi, Qingming Huang, and Qi Tian. Centernet: Keypoint triplets for object detection. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pages 6569–6578, 2019.

[12] 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.

[13] Yaroslav Ganin and Victor Lempitsky. Unsupervised domain adaptation by backpropagation. In International conference on machine learning, pages 1180–1189. PMLR, 2015.

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

[15] 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.

[16] 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.

[17] 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.

[18] 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.

[19] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 4700–4708, 2017.

[20] Jiaxing Huang, Dayan Guan, Aoran Xiao, and Shijian Lu. Cross-view regularization for domain adaptive panoptic segmentation. arXiv preprint arXiv:2103.02584, 2021.

[21] Jiaxing Huang, Dayan Guan, Aoran Xiao, and Shijian Lu. Fsdr: Frequency space domain randomization for domain generalization. arXiv preprint arXiv:2103.02370, 2021.

[22] Jiaxing Huang, Shijian Lu, Dayan Guan, and Xiaobing Zhang. Contextual-relation consistent domain adaptation for semantic segmentation. In European Conference on Computer Vision, pages 705–722. Springer, 2020.

[23] Naoto Inoue, Ryosuke Furuta, Toshikiko 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.

[24] 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.

[25] 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.

[26] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. Advances in neural information processing systems, 25:1097–1105, 2012.
[27] Hei Law and Jia Deng. Cornernet: Detecting objects as paired keypoints. In Proceedings of the European conference on computer vision (ECCV), pages 734–750, 2018. 1, 2

[28] 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, 7, 12, 13

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

[30] 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

[31] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In Proceedings of the IEEE international conference on computer vision, pages 2980–2988, 2017. 1

[32] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, C, and Lawrene Zitnick. Microsoft coco: Common objects in context. In European conference on computer vision, pages 740–755, 2014. 1

[33] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander Berg. Ssd: Single shot multibox detector. In European conference on computer vision, pages 21–37. Springer, 2016. 1, 2

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

[35] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 779–788, 2016. 1, 2

[36] Joseph Redmon and Ali Farhadi. Yolo9000: better, faster, stronger. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 7263–7271, 2017. 2

[37] Joseph Redmon and Ali Farhadi. Yolov3: An incremental improvement. arXiv preprint arXiv:1804.02767, 2018. 2

[38] Shaqing 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, 2, 6, 7

[39] 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. 1, 2, 5, 6, 7

[40] 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. 1

[41] Kuniaki Saito, Shohei Yamamoto, Yoshitaka Ushiku, and Tatsuya Harada. Open set domain adaptation by backpropagation. In The European Conference on Computer Vision (ECCV), September 2018. 1

[42] 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. 1, 5

[43] 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

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

[45] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556, 2014. 2

[46] Yi-Hsuan Tsai, Wei-Chih Hung, Samuel Schultze, Kihyuk Sohn, Ming-Hsuan Yang, and Manmohan Chandraker. Learning to adapt structured output space for semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 7472–7481, 2018. 1

[47] 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

[48] 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, 8

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

[50] Chang-Dong Xu, Xing-Ran Zhao, Xin Jin, and Xi-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. 1, 2, 6, 7

[51] 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. 1

[52] Fuxun Yu, Di Wang, Yinpeng Chen, Nikolaos Karaiskakis, 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

[53] Gongjie Zhang, Kaiwen Cui, Rongliang Wu, Shijian Lu, and Yonghong Tian. Ppdpet: Efficient few-shot detection without forgetting via plug-and-play sub-networks. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pages 3823–3832, 2021. 1
[54] Gongjie Zhang, Shijian Lu, and Wei Zhang. Cad-net: A context-aware detection network for objects in remote sensing imagery. *IEEE Transactions on Geoscience and Remote Sensing*, 57(12):10015–10024, 2019. 1

[55] Gongjie Zhang, Zhipeng Luo, Kaiwen Cui, and Shijian Lu. Meta-detr: Few-shot object detection via unified image-level meta-learning. *arXiv preprint arXiv:2103.11731*, 2021. 1

[56] 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, 7

[57] 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

[58] 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, 6, 7, 8, 12, 13

[59] Chenfan Zhuang, Xintong Han, Weilin Huang, and Matthew Scott. ifan: Image-instance full alignment networks for adaptive object detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 13122–13129, 2020. 1
6. Appendix

This section provides more qualitative results of our proposed method which are omitted in the main paper due to space limitation.

6.1. Qualitative Detection Results

We present more qualitative experimental results and compare the proposed DA-DETR with standard DETR [58] and state-of-the-art method [28]. Fig. 5 shows experimental results over the synthetic-to-real adaptation task SIM 10k → Cityscapes. It can be seen that the standard DETR produces a number of false detections (e.g. in the sample images in the third and fourth rows) due to domain gaps. SAP [28] generates more precise bounding boxes but misses some small object (e.g. in the sample images in the third and fifth rows). The proposed DA-DETR adapts from synthetic to real well and can detect more small objects with less false alarms as illustrated.

6.2. Visualization of Attention

We further visualize the attention that is generated by the proposed hybrid attention module (HAM). Fig. 6 shows several sample images (from the validation dataset of Cityscapes) and the corresponding attention maps, where the model is trained over the synthetic-to-real adaptation task SIM 10k → Cityscapes. It can be observed that the HAM-predicted attention guides the network to focus on informative and hard-aligned regions accurately.
Figure 5. Qualitative illustration of domain adaptive detection over the task SIM 10k→Cityscapes: The proposed DA-DETR exploits hybrid attention which adapts well from synthetic to real scenes and detects most objects in Target Images (from Cityscapes) accurately. DETR [58] and SAP [28] do not take full advantage of different types of attention which leads to sub-optimal cross-domain alignment and object detection performance.
Figure 6. Visualization of attention generated by the proposed hybrid attention module (HAM): For eighteen sample images from the validation set of Cityscapes as shown in columns 1 and 3, columns 2 and 4 show the corresponding attention that is predicted by the proposed hybrid attention module (HAM). It can be observed that the HAM-predicted attention detects informative and hard-aligned regions accurately.