Mean Teacher DETR with Masked Feature Alignment:
A Robust Domain Adaptive Detection Transformer Framework

Weixi Weng, Chun Yuan

Tsinghua Shenzhen International Graduate School, Tsinghua University
wengwx22@mails.tsinghua.edu.cn, yuanc@sz.tsinghua.edu.cn

Abstract

Unsupervised domain adaptive object detection (UDAOD) research on Detection Transformer (DETR) mainly focuses on feature alignment and existing methods can be divided into two kinds, each of which has its unresolved issues. One-stage feature alignment methods can easily lead to performance fluctuation and training stagnation. Two-stage feature alignment method based on mean teacher comprises a pretraining stage followed by a self-training stage, each facing problems in obtaining a reliable pretrained model and achieving consistent performance gains. Methods mentioned above have not yet explored how to utilize the third related domain such as the target-like domain to assist adaptation. To address these issues, we propose a two-stage framework named MTM, *i.e.* **M**ean **T**eacher-DETR with **M**asked Feature Alignment. In the pretraining stage, we utilize labeled target-like images produced by image style transfer to avoid performance fluctuation. In the self-training stage, we leverage unlabeled target images by pseudo labels based on mean teacher and propose a module called Object Queries Knowledge Transfer (OQKT) to ensure consistent performance gains of the student model. Most importantly, we propose masked feature alignment methods including Masked Domain Query-based Feature Alignment (MDQFA) and Masked Token-Wise Feature Alignment (MTWFA) to alleviate domain shift in a more robust way, which not only prevent training stagnation and lead to a robust pretrained model in the pretraining stage but also enhance the model's target performance in the self-training stage. Experiments on three challenging scenarios and a theoretical analysis verify the effectiveness of MTM.

Introduction

Object detection is always recognized as an important task in the field of computer vision. Recently, Detection Transformer (DETR) (Carion et al. 2020) redefines object prediction by departing from the traditional anchor-based methodology (Girshick 2015) and embracing the concept of object queries which serve as learnable embeddings interacting with the image features, allowing the model to predict object classes and bounding box coordinates. While DETR has demonstrated remarkable performance on various datasets, its application to real-world environments still presents challenges, particularly when training data and testing data are

collected from different distributions, *i.e.* domain shift.

In order to enable object detectors trained on a labeled source domain to be effectively deployed on a completely unlabeled target domain, unsupervised domain adaptive object detection(UDAOD) emerged as a solution. UDAOD research on DETR is gradually gaining increasing attention, with particular emphasis on feature alignment (Wang et al. 2021; Gong et al. 2022; Yu et al. 2022; Zhang et al. 2023). Feature alignment is commonly implemented through adversarial training, with the objective of extracting domain-invariant global-level, local-level and instance-level features from source domain and target domain, thus reducing domain shift. Existing methods on DETR can be categorized into one-stage feature alignment methods (Wang et al. 2021; Gong et al. 2022; Zhang et al. 2023) and two-stage feature alignment method based on mean teacher (Yu et al. 2022). We deeply studied the above two methods and identified several inherent issues, which will be elaborated below.

One-stage feature alignment methods (Xu et al. 2020a; Gong et al. 2022; Zhang et al. 2023) generally train DETR from scratch with labeled source images and unlabeled target images, conducting feature alignment on the output of the backbone, encoder, and decoder. However, training in this manner has inherent issues: (I1) performance fluctuation: the lack of labels for target images restricts the model's ability to extract features from target images, and conducting feature alignment with low-quality target features easily results in performance fluctuation. Furthermore, it's very likely to end up with an underperforming model as shown by the red line in Figure 1; (I2) training stagnation: at the early training stage, if there is a significant difference in the distribution between the two types of training data, domain discriminators converge quickly and become proficient at distinguishing features in the later period of training. However, the backbone struggles to produce domain-invariant features at this point, which results in marginal performance improvement as shown by the yellow line in Figure 1.

Two-stage feature alignment method (Yu et al. 2022) based on mean teacher (Sohn et al. 2020; Xu et al. 2021) aims to leverage unlabeled target images by self-training. Take MTTrans (Yu et al. 2022) for example, it comprises a pretraining stage followed by a self-training stage. It firstly pretrains a model merely on labeled source images and only conducts feature alignment in the self-training stage, shar-
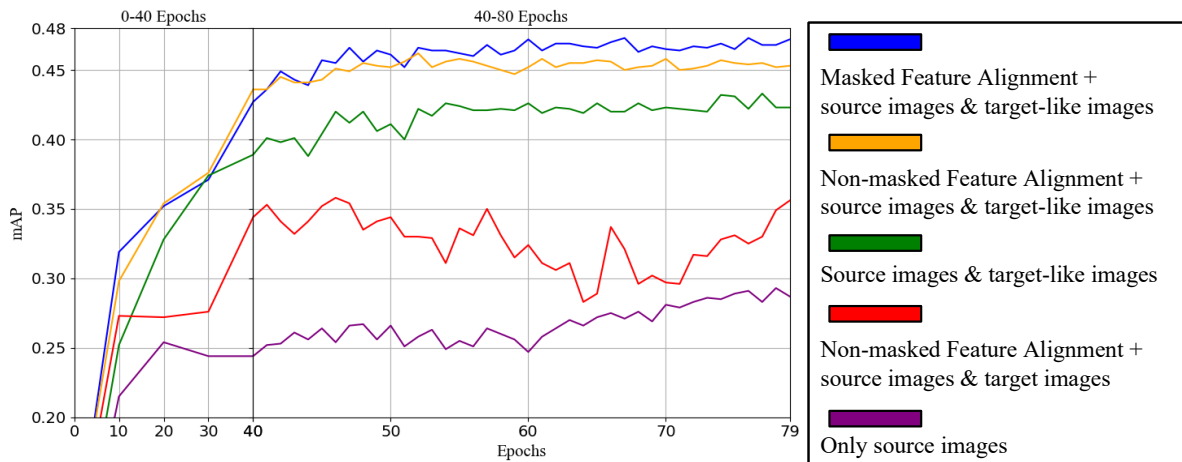
Figure 1: We train Deformable DETR from scratch with different pretraining strategies in the weather adaptation scenario. It turns out that pretraining with masked feature alignment on source images and target-like images brings the best performance.

ing the same issues with the one-stage methods. Although MTTrans significantly improves target performance through mean teacher, it also faces other issues: (I3) unreliable pretrained model: its pretrained model is obtained by training on labeled source images without any feature alignment, resulting in the teacher model failing to generate accurate pseudo-labels; (I4) unstable performance gains: it uses the same the object queries between the teacher and the student model in the self-training stage, resulting in the student model struggling to attain consistent performance gains.

Furthermore, the methods mentioned above only consider adapting between two domains, while how to use the third related domain such as a target-like domain generated by image style transfer to assist adaptation is still unexplored.

Based on the facts above, we propose MTM, *i.e.* **M**ean **T**eacher-DETR with **M**asked Feature Alignment consisting of a pretraining stage and a self-training stage. In the pretraining stage, we utilize labeled source images and labeled target-like images produced by cycleGAN (Zhu et al. 2017) instead of unlabeld target images to avoid the problem of performance fluctuation (I1). In the self-training stage, we adopt the mean teacher framework to leverage target images by pseudo labels. We further propose Object Queries Knowledge Transfer (OQKT) which enhances the semantic information of the student model's object queries by multi-head attention to prompt consistent performance gains of the student model (I4). Most importantly, we propose masked feature alignment methods including Masked Domain Query-based Feature Alignment (MDQFA) and Masked Token-Wise Feature Alignment (MTWFA) to alleviate domain shift in a more robust way, which not only prevent training stagnation (I2) and lead to robust pretrained model (I3) in the pretraining stage but also enhance the model's final target performance in the self-training stage. The contributions of this paper are as follows:

- Through experiments we find that utilizing labeled target-like images produced by CycleGAN (Zhu et al.

2017) to participate in training DETR from scratch avoids performance fluctuation.

- We propose Object Queries Knowledge Transfer (OQKT) based on mean teacher DETR to guarantee consistent performance gains of the student model.

- Most importantly, we propose masked feature alignment methods including Masked Domain Query-based Feature Alignment (MDQFA) and Masked Token-Wise Feature Alignment (MTWFA) to reduce domain shift in a more robust way. They benefit both stages of MTM.

- Extensive experiments on three challenging domain adaptation scenarios have demonstrated that MTM has outperformed existing state-of-the-art (SOTA) methods in this field. A theoretical analysis is also presented to verify the effectiveness of MTM.

Related Work

Detection Transformer

Object detection is a challenging task in the field of computer vision. Recently, the Detection Transformer(DETR) has garnered significant attention by presenting a novel object detection pipeline based on Transformer. Deformable DETR (Zhu et al. 2021), as an important advancement of DETR, introduces deformable attention modules that only attend to a small set of key sampling points around a reference. Deformable DETR achieves better performance than DETR with much less convergence time and computational requirements. Deformable DETR has several variants and the two-stage variant achieves the best performance. In the two-stage Deformable DETR, the encoder generates object queries from multi-scale features extracted by the CNN backbone, while the decoder refines object queries and employs them for predicting object classes and generating bounding box coordinates. Following (Wang et al. 2021), we mainly investigate UDAOD on Deformable DETR, specifically the two-stage Deformable DETR.

Domain Adaptive Object Detection

Extensive research has been conducted on unsupervised domain adaptive object detection (UDAOD) based on different architectures. Proposed solutions can be categorized into three mainstreams: feature alignment, mean teacher and domain transfer. These kinds of solutions often work together to form a more robust framework. However, research on DETR so far has mainly focused on feature alignment.

DAF (Chen et al. 2018) is the first work to apply feature alignment on Faster RCNN by means of adversarial training, performing feature alignment at the image-level and instance-level respectively. Recently, there has been a continuous surge of research based on DETR. SFA (Wang et al. 2021) makes a pioneering attempt to conduct sequence feature alignment on different levels based on DETR. O²Net (Gong et al. 2022) introduces an Object-Aware Alignment module to align the multi-scale features of the CNN backbone. DA-DETR (Zhang et al. 2023) proposes a CNN-Transformer Blender which fuses the output of the CNN backbone and the encoder to better align them.

MTOR (Cai et al. 2019) firstly exploits mean teacher on Faster RCNN, building a consistency loss between the teacher and the student model to learn more about the object relation. MTTrans (Yu et al. 2022) firstly applies mean teacher on DETR, proposing multi-level feature alignment to improve the quality of pseudo labels.

Methods of domain transfer are normally built on CycleGAN (Zhu et al. 2017) which conducts image style transfer between domains. Given source images and target images, CycleGAN generates target-like images by transferring source images into target style, as well as source-like images. UMT (Deng et al. 2021) additionally utilizes source-like and target-like images for training, and proposes several strategies on mean teacher to address model bias. AWADA (Menke, Wenzel, and Schwung 2022) uses the proposals generated by the object detector to aid in more efficient image style transfer. While domain transfer has been extensively studied on other object detectors (Ren et al. 2015), to the best of our knowledge, how to use it to assist domain adaptation on DETR remains unexplored.

Methods

This section introduces MTM consisting of a robust pretraining stage and a performance-enhancing self-training stage.

MTM Framework Overview

In the pretraining stage, we utilize labeled source images and labeled target-like images for training. A target-like image produced by CycleGAN (Zhu et al. 2017) retains the content of the source image while exhibiting the style of target images, thereby sharing identical annotations with its corresponding source images. Pretraining with labeled target-like images not only avoids performance fluctuation (I1) as shown by the green line in Figure 1 but also contributes to obtaining a robust pretrained model (I3).

In the self-training stage, we utilize labeled source images and unlabeled target images for training. Given the absence of target images during pretraining, we leverage them by pseudo labels based on mean teacher in this stage. We further propose Object Queries Knowledge Transfer (OQKT) based on mean teacher DETR to ensure consistent performance gains of the student model (I4).

Most importantly, we propose masked feature alignment methods including Masked Domain Query-based Feature Alignment (MDQFA) and Masked Token-Wise Feature Alignment (MTWFA) which not only prevent training stagnation (I2) and help to obtain a robust pretrained model (I3) in the pretraining stage but also contribute to enhancing the model's final target performance in the self-training stage.

Object Queries Knowledge Transfer

In the mean teacher framework, the teacher model takes in weakly augmented target images as input while the student model processes strongly augmented counterparts. As they undergo different data augmentation strategies, it is highly likely to result in significantly different object queries between them. Therefore, enforcing the student to use the same object queries as the teacher (Yu et al. 2022) tends to restrain consistent performance gains of the student. However, we can't ignore that the teacher model is likely to generate valuable object queries from weakly augmented target images. In order to transfer the knowledge within the object queries from the teacher model to the student model and address the issue of unstable performance gains (I4), we propose a simple yet effective module named **O**bject **Q**ueries **K**nowledge **T**ransfer(OQKT).

QE_t represents query embeddings of the teacher model's object queries and PE_t is the corresponding position embeddings. The teacher model's object queries are obtained by adding QE_t and PE_t together. In a similar way, QE_s and PE_s are defined. We further define:

$$Query = QE_s + PE_s$$
$$Key = QE_t + PE_t \qquad (1)$$
$$Value = QE_t$$

The knowledge transfer mechanism of this module is based on the multi-head attention mechanism in (Vaswani et al. 2017), which is shown as follows:

$$\text{ATTENTION}(Q, K, V) = \text{softmax}(\frac{QK^T}{\sqrt{d_k}}V) \qquad (2)$$

$$\text{MULTIHEAD}(Q, K, V) = \text{Concat}(head_1, \cdots head_n)W^O$$
$$where\ head_i = \text{ATTENTION}(QW_i^Q, KW_i^K, VW_i^V) \qquad (3)$$

Where W_i^Q, W_i^K, W_i^V and W_i^O are weight matrices and $n = 16, d_k = 16$ in OQKT. The teacher's object queries and the student's object queries interact via the multi-head attention mechanism to obtain additional features. The additional features are multiplied by a parameter and added to the student's query embeddings, enhancing the semantic information of the student's object queries for better prediction. QE_S is updated as follows:

$$QE_S = QE_S + \alpha \times \text{MULTIHEAD}(Query, Key, Value) \quad (4)$$

In the self-training stage, with the increasing of epochs, α decreases linearly from 1 to 0, promoting the student model to generate high-quality object queries by itself.
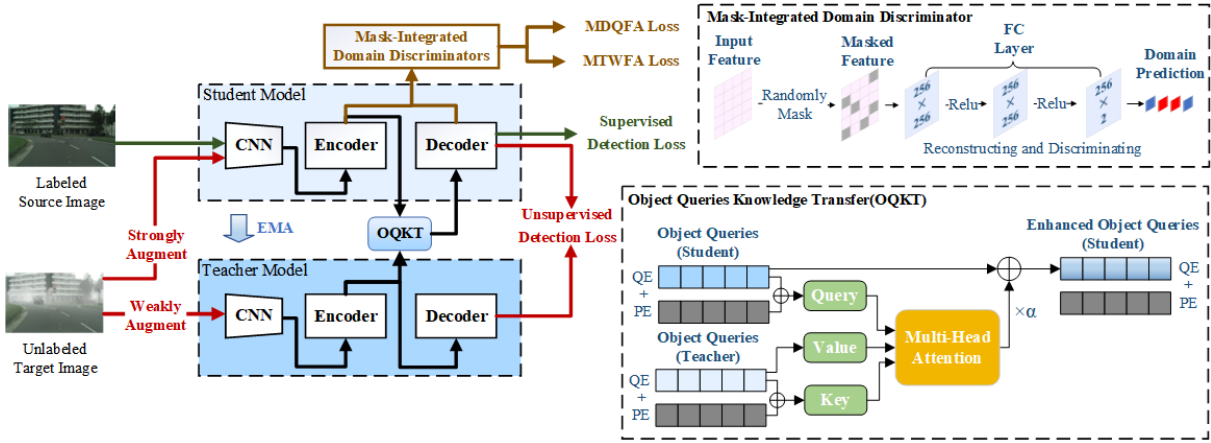
Figure 2: MTM framework in the self-training stage. In the mean teacher structure, the student model is updated by back-propagation while the teacher model is updated by the Exponential Moving Average (EMA) of the student model. OQKT enhances the student's object queries with the teacher's object queries by the multi-head attention mechanism. Masked feature alignment is also conducted in this stage to alleviate the domain shift between source data and target data.

Masked Feature Alignment

Proposed by (Wang et al. 2021), Domain Query-based Feature Alignment (DQFA) alleviates global-level domain shift from image style, weather *etc.*, while Token-Wise Feature Alignment (TWFA) deals with local-level and instance-level domain shift caused by object appearance, scale, texture *etc*. However, when conducting them on source images and target-like images in the pretraining stage, training stagnation is observed as shown by the yellow line in Figure 1.

To prevent training stagnation and elevate the model's target performance, we innovatively propose masked feature alignment methods including Masked Domain Query-based Feature Alignment (MDQFA) and Masked Token-Wise Feature Alignment (MTWFA). MDQFA and MTWFA randomly mask sequence features before feeding them into domain discriminators, making it harder to classify their domains. Our masked feature alignment methods not only hinder the domain discriminators from converging too quickly thus preventing training stagnation, but also improve the domain discriminators' robustness through a diverse range of masked sequence features, consequently further elevating the model's target performance. Details of MDQFA and MTWFA are described below. The domain discriminators' structures for both masked feature alignment methods exhibit identical architectures, as depicted in Figure 2.

Masked Domain Query-based Feature Alignment On the encoder side, a domain query $q^{enc} \in R^{1 \times C}$ is concatenated with the multi-scale feature extracted from the CNN backbone to form the input for the encoder:

$$Z_0 = [q^{enc}, z_1, z_2, ..., z_{N_{enc}}] + E_{pos} + E_{level} \quad (5)$$

N_{enc} stands for the length of the encoder input embeddings. E_{pos} and E_{level} represents corresponding position embeddings and feature level embeddings. The sequence features derived from the l-th layer are denoted as Z_l. During the encoding process, the domain query gathers domain-specific features from the sequence feature. We further set a random mask which can be formulated as:

$$M^{\text{DQFA}}_{enc_{l,i}} = \begin{cases} 0, & \text{if } R_{l,i} < \theta_{Mask} \\ 1, & \text{Otherwise} \end{cases} \quad (6)$$

where $l = 1...L_{enc}$ indexes the layer of the encoder, $i \in [0, C)$ represents the element coordinate of domain query, $R_{l,i}$ is a random floating-point number that falls within the range of 0 to 1 and θ_{Mask} is a hyperparameter controlling the masking rate. Then we compute the element-wise product of the domain query and the mask to get the masked domain query $Z^M_{l,0}$ that only contains partial global domain information. The domain discriminator D^{MDQFA}_{enc} needs to determine which domain the masked domain query belongs to. We utilize the binary cross-entropy loss as the domain classification loss. The whole procedure is formulated as below:

$$Z^M_{l,0} = M^{\text{DQFA}}_{enc_l} \odot Z_{l,0} \quad (7)$$

$$\mathcal{L}^{\text{MDQFA}}_{enc_l} = -\big[dlogD^{\text{MDQFA}}_{enc}(Z^M_{l,0}) \\ + (1-d)log\big(1 - D^{\text{MDQFA}}_{enc}(Z^M_{l,0})\big)\big] \quad (8)$$

d represents the domain label with 0 denoting the source domain and 1 denoting the target or target-like domain.

On the decoder side, we concatenate a domain query $q^{dec} \in R^{1 \times C}$ with the object queries to form the input:

$$Q_0 = [q^{dec}; q_1, q_2, ..., q_{N_{dec}}] + E'_{pos} \quad (9)$$

N_{dec} stands for the length of the decoder input embeddings. E'_{pos} represents position embeddings. The refined object queries derived from the l-th layer are denoted as Q_l. The generation of the random mask of the domain queries in the decoder $M^{\text{DQFA}}_{dec_l}$ follows a similar procedure as equation 7. We feed the masked domain query into the domain discriminator D^{MDQFA}_{dec} and calculate the binary cross-entropy loss.

Masked Token-Wise Feature Alignment On the encoder side, each token embedding in the sequence feature stands for a particular-scale feature of the image and aggregates local-level information from nearby key points. We set a random mask for the whole sequence feature as follows:

$$M_{enc_{l,i,j}}^{\text{TWFA}} = \begin{cases} 0, & \text{if } R_{l,i,j} < \eta \cdot \theta_{Mask} \\ 1, & \text{Otherwise} \end{cases} \quad (10)$$

where $l = 1...L_{dec}$ indexes the layers of the encoder, $i \in [1, N_{enc}]$ and $j \in [0, C)$ respectively represent the horizontal and vertical coordinates of the sequence feature. Since domain discriminators of MTWFA deal with more diverse local-level features which are more difficult to classify their domains, we set the hyperparameter η to a decimal between 0 and 1 indicating that MTWFA will use a smaller mask threshold compared with MDQFA. Then we compute the element-wise product of the sequence feature and the corresponding mask as the masked sequence feature, which is fed into the domain discriminator D_{enc}^{MTWFA}. We adopt the binary cross-entropy loss as the domain classification loss. The whole procedure is formulated as below:

$$Z_l^M = M_{enc_l}^{\text{TWFA}} \odot Z_l \quad (11)$$

$$\mathcal{L}_{enc_l}^{\text{MTWFA}} = -\frac{1}{N_{enc}} \sum_{i=1}^{N_{enc}} \left[d \log D_{enc}^{\text{MTWFA}}(Z_{l,i}^M) \right.$$
$$\left. + (1-d) \log \left(1 - D_{enc}^{\text{MTWFA}}(Z_{l,i}^M)\right) \right] \quad (12)$$

d signifies the domain label, where 0 denotes the source domain, and 1 denotes the target or target-like domain.

On the decoder side, each object query that corresponds to a predicted object's class and position aggregates instance-level features in the decoding process. Similarly, the sequence features of the decoder are randomly masked and then fed into the domain discriminator D_{dec}^{MTWFA} to calculate the corresponding binary cross-entropy loss.

Overall Training Strategy
Our MTM framework consists of two training stages. In the pretraining stage, we aim to train a robust pretrained model, which will serve as the teacher model later. The loss function of this stage L_{pre} combines the object detection loss and the feature alignment loss, which is defined as follows:

$$\mathcal{L}_{pre} = \mathcal{L}_{det}(I_s, B_s) + \mathcal{L}_{det}(I_{tl}, B_s) - \mathcal{L}_{adv}(I_s) - \mathcal{L}_{adv}(I_{tl}) \quad (13)$$

Where I_s and I_{tl} represent source images and target-like images, and they share the same annotations B_s. The feature alignment loss $\mathcal{L}_{adv}$ consists of four parts, which are MDQFA loss on the encoder, MTWFA loss on the encoder, MDQFA loss on the decoder and MTWFA on the decoder:

$$\mathcal{L}_{adv} = \sum_{l=1}^{L_{enc}} (\lambda_{\text{MDQFA}} \cdot \mathcal{L}_{enc_l}^{\text{MDQFA}} + \lambda_{\text{MTWFA}} \cdot \mathcal{L}_{enc_l}^{\text{MTWFA}}) +$$
$$\sum_{l=1}^{L_{dec}} (\lambda_{\text{MDQFA}} \cdot \mathcal{L}_{dec_l}^{\text{MDQFA}} + \lambda_{\text{MTWFA}} \cdot \mathcal{L}_{dec_l}^{\text{MTWFA}}) \quad (14)$$

After obtaining a robust pretrained model in the pretraining stage, we then proceed to conduct self-training. The loss function of this stage L_{st} can be formulated as:

$$\mathcal{L}_{st} = \mathcal{L}_{det}(I_s, B_s) + \mathcal{L}_{det}(I_t, \hat{B}_t) - \mathcal{L}_{adv}(I_s) - \mathcal{L}_{adv}(I_t) \quad (15)$$

Where I_t represents target images and $\hat{B}_t$ refers to pseudo labels produced by the teacher model for I_t. To summarize, the overall training objective of MTM is defined as:

$$\min_G \max_D \mathcal{L}_{det}(G) - \mathcal{L}_{adv}(G, D) \quad (16)$$

where G is the object detector and D represents the domain discriminators.

Experiment

In this section, we conduct extensive experiments to testify our contributions: (1) pretraining with labeled target-like images avoids performance fluctuation (I1) and improves target performance; (2) OQKT helps the student model earn consistent performance gains in the self-training stage (I4); (3) our masked feature alignment methods including MDQFA and MTWFA prevent training stagnation (I2) and provide significant performance improvements in both stages.

Datasets and Settings
Datasets We evaluate MTM on three challenging domain adaptation scenarios. *i.e.* weather adaptation, scene adaptation, and synthetic to real adaptation. A detailed introduction of the three scenarios is as below:

- Weather Adaptation: Cityscapes (Cordts et al. 2016) is a landscape dataset containing 2975 training and 500 validation images. Foggy Cityscapes (Sakaridis, Dai, and Van Gool 2018) is generated from Cityscapes by a fog synthesis algorithm. In this scenario, Cityscapes is the source dataset and Foggy Cityscapes is the target dataset.

- Scene Adaptation: In this scenario, Cityscapes also serves as the source dataset. BDD100K (Yu et al. 2020) is an autonomous driving dataset consisting of 100k HD video clips. We utilize the *daytime* subset of BDD100K which contains 36728 training images and 5258 validation images as the target dataset in this scenario.

- Synthetic to Real Adaptation: Sim10K (Johnson-Roberson et al. 2017) is generated by the Grand Theft Auto game engine, containing 10,000 synthetic game images. In this scenario, Sim10K serves as the source dataset and Cityscapes serves as the target dataset.

Comparative Benchmarks We compare MTM with five baselines built on Deformable DETR to validate the effectiveness of our proposed framework: Deformable DETR only trained on source data (Zhu et al. 2021), SFA (Wang et al. 2021), O^{2}Net (Gong et al. 2022), MTTrans (Yu et al. 2022) and DA-DETR (Zhang et al. 2023).

Evaluation Metric We report the Average Precision on the car category with a threshold of 0.50 in the synthetic to real adaptation scenario. We adopt the mean Average Precision (mAP) of a threshold of 0.50 in the other two scenarios.

Implementation Details We train CycleGAN (Zhu et al. 2017) for 100 epochs with a batch size of 8 and a learning rate of 2×10^{-4} which linearly decreases to 0 after 50 epochs. We adopt an ImageNet (Deng et al. 2009)-pretrained ResNet50 network as the CNN backbone. In the pretraining

Method	person	rider	car	truck	bus	train	motorcycle	bicycle	mAP
Deformable DETR(source)	38.0	38.7	45.3	16.3	26.7	4.2	22.9	36.7	28.5
SFA(Wang et al. 2021)	46.5	48.6	62.6	25.1	46.2	29.4	28.3	44.0	41.3
MTTrans(Yu et al. 2022)	47.7	49.9	65.2	25.8	45.9	33.8	32.6	46.5	43.4
DA-DETR(Zhang et al. 2023)	49.9	50.0	63.1	25.8	45.9	33.8	32.6	46.5	43.5
O^{2}Net(Gong et al. 2022)	48.7	51.5	63.6	31.1	47.6	**47.8**	38.0	45.9	46.8
MTM(ours)	**51.0**	**53.4**	**67.2**	**37.2**	**54.4**	41.6	**38.4**	**47.7**	**48.9**

Table 1: Results(%) in the weather adaptation scenario, *i.e.* Cityscapes → Foggy Cityscapes.

Method	person	rider	car	truck	bus	motorcycle	bicycle	mAP
Deformable DETR(source)	38.9	26.7	55.2	15.7	19.7	10.8	16.2	26.2
SFA(Wang et al. 2021)	40.4	27.6	57.5	19.1	23.4	15.4	19.2	28.9
O^{2}Net(Gong et al. 2022)	40.4	31.2	58.6	20.4	25.0	14.9	22.7	30.5
MTTrans(Yu et al. 2022)	44.1	30.1	61.5	25.1	26.9	17.7	23.0	32.6
MTM(ours)	**53.7**	**35.1**	**68.8**	**23.0**	**28.8**	**23.8**	**28.0**	**37.3**

Table 2: Results(%) in the scene adaptation scenario, *i.e.* Cityscapes → BDD100K.

Method	AP on car
Deformable DETR(source)	47.4
SFA(Wang et al. 2021)	52.6
O^{2}Net(Gong et al. 2022)	54.1
DA-DETR(Zhang et al. 2023)	54.7
MTTrans(Yu et al. 2022)	57.9
MTM(ours)	**58.1**

Table 3: Results(%) in the synthetic to real adaptation scenario, *i.e.* Sim10k → Cityscapes.

stage of 80 epochs, we use an Adam optimizer with an initial learning rate 2×10^{-4} decayed by 0.1 every 40 epochs. In the self-training stage of 20 epochs, the learning rate is 2×10^{-6}. The filtering threshold of pseudo labels is 0.50. Following (Wang et al. 2021), in the weather adaptation scenario, λ_{MTWFA} is 1, and λ_{MDQFA} is 0.1. In other scenarios, they are set to 0.01 and 0.001 respectively. θ_{Mask} and η are set to 0.40 and 0.50. We use one 24GB GeForce RTX 3090 GPU in all experiments. Each batch includes 1 image from the source domain and 1 image from either the target-like domain or the target domain.

Comparative Study

The evaluations of our MTM are conducted on three domain adaptation scenarios. In the weather adaptation scenario (Table 1), MTM outperforms the SOTA by 2.1 mAP. In the scene adaptation scenario (Table 2), MTM shows a significant performance improvement of 4.7 mAP over the current SOTA method. In the synthetic to real adaptation scenario (Table 3), MTM demonstrates a performance improvement of 0.2 mAP compared to MTTrans (Yu et al. 2022), but still has a large improvement compared to the other baselines.

Ablation Studies

Given that our framework is made up of two stages, the ablation studies are conducted in each stage in the weather adaptation scenario to prove the effectiveness of each component. Results of the ablation studies are presented in Table 4-6.

Pretraining Stage Several observations can be drawn from Table 4: (1) pretraining with target-like images significantly enhances the model's target performance, resulting in a 13.8 mAP improvement; (2) MDQFA and MTWFA further improve performance compared to their non-masked counterparts. MDQFA and MTWFA further improve the model's target performance by 0.8 mAP and 0.9 mAP compared to their non-masked counterpart respectively. Combining both of them culminates in a 1.9 mAP improvement over their non-masked counterparts combination.

As depicted in Figure 1, we can observe that: (1) training with target-like images indeed avoids performance fluctuation (I1); (2) when conducting non-masked feature alignment on source images and target-like images, target performance plateaus after 40 epochs, indicating training stagnation (I2). Yet, masked feature alignment prevents this issue and maintains performance growth even after 40 epochs.

In Table 5, we present the target performance of pretrained models under different mask settings controlled by hyperparameters θ_{Mask} and η. The best pretraining performance is attained when $\theta_{Mask} = 0.40$ and $\eta = 0.50$. From Table 5, we find that setting the mask threshold θ_{Mask} of MDQFA between 0.30 and 0.50, and the mask threshold $\eta \cdot \theta_{Mask}$ of MTWFA between 0.10 and 0.30, further enhances the target performance of the pretrained model. The disparate optimal threshold ranges of the two methods are likely due to the fact that they handle features of different levels. MDQFA deals with global domain queries, so its domain discriminators are capable of distinguishing domains even when a significant portion of the information is masked. However, MTWFA handles local-level and instance-level sequence features, thus it will easily cause confusion to the domain discriminators of MTWFA if too much information is dropped from the mask.

Self-training Stage Table 6 reveals notable observations as below: (1) MT helps increase target performance; (2) OQKT helps MT framework further improve target performance. With a pretrained model of 47.2 mAP, MT with OQKT outperforms standard MT by 0.2 mAP. MT achieves

Method	Target-like	MTWFA	MDQFA	mAP
D-DETR				28.5
	✓			42.3
	✓	◯		42.7
	✓	✓		43.6
Proposed	✓		◯	44.9
	✓		✓	45.7
	✓	◯	◯	45.3
	✓	✓	✓	**47.2**

Table 4: Ablation study of the pretraining stage in the weather adaptation scenario. D-DETR stands for Deformable-DETR trained on source data. ◯ refers to adopting the non-masked feature alignment method.

θ_{Mask}		0.20	0.30	0.40	0.50	0.60	0.70	0.80
	1/3	45.3	45.5	46.1	**46.8**	45.2	45.0	44.5
η	1/2	45.6	46.2	**47.2**	45.8	44.9	44.3	44.2
	1	45.6	45.8	**46.6**	45.5	44.9	44.6	43.4

Table 5: Performance of pretrained models in the weather adaptation scenario under different mask settings.

Method	MT	OQKT	MTWFA	MDQFA	mAP
D-DETR					28.5
	✓				35.8
	✓	✓			38.3
					47.2
	✓				47.7
Proposed	✓	✓			47.9
	✓	✓		✓	48.0
	✓	✓	✓		48.1
	✓	✓		✓	48.4
	✓	✓	✓	✓	**48.9**

Table 6: Ablation study of the self-training stage in the weather adaptation scenario. MT refers to utilizing the mean teacher framework for self-training.

a further performance gain of 2.5 mAP if incorporated with OQKT; (3) MDQFA and MTWFA improve target performance by 0.1 mAP and 0.2 mAP respectively, and their combination brings a performance gain of 0.4 mAP.

Theoretical Analysis

This section analyzes our framework from a theoretical aspect based on (Blitzer et al. 2007; Ben-David et al. 2010).

Let $\mathcal{H}$ be a hypothesis space of VC dimension d. For each $i \in \{1, ..., N\}$, let S_i be a labeled sample of size $\beta_j m$ generated by drawing $\beta_j m$ points from domain $\mathcal{D}_i$ and labeling them according to f_i, and α_i represents domain weight of $\mathcal{D}_i$. N is set to 2 in our work because we only use two types of labeled images including source images and target-like images. If $\hat{h} \in \mathcal{H}$ is an empirical hypothesis of the empirical α-weighted error $\hat{\epsilon}_\alpha(h)$ on these multi-source samples, we have the following theorem to bound the target error of the empirical hypothesis $\hat{h}$:

Theorem 1 *For any $\delta \in (0, 1)$, with probability $1 - \delta$,*

$$\epsilon_T(\hat{h}) \leq \hat{\epsilon}_\alpha(\hat{h}) + \frac{1}{2} d_{\mathcal{H}\Delta\mathcal{H}}(D_\alpha, D_T) + \gamma_\alpha$$
$$+ \sqrt{(\sum_{j=1}^2 \frac{\alpha_j^2}{\beta_j})(\frac{dln(2m) - ln(\delta)}{2m})} \quad (17)$$

where $\gamma_\alpha = min_h\{\epsilon_T(h) + \epsilon_\alpha(h)\}$ represents the error of the joint ideal hypothesis that is correlated with the ability of the detector and the discriminability of features. $d_{\mathcal{H}\Delta\mathcal{H}}(D_\alpha, D_T)$ is the domain divergence that is associated with the domain-invariance of features.

In Theorem 1, the target error $\epsilon_T(\hat{h})$ can be bounded by four factors: (1) the empirical α-weighted error on the multi-source samples $\hat{\epsilon}_\alpha(\hat{h})$, which can be minimized by applying supervised loss on the multi-source samples; (2) the domain divergence between the multi-source and the target domain $d_{\mathcal{H}\Delta\mathcal{H}}(D_\alpha, D_T)$. Our masked feature alignment methods alleviate domain shift to minimize the domain divergence; (3) the error of the joint ideal hypothesis γ_α. Both our pretraining stage and self-training stage help enhance the model's ability to reduce this value; (4) a complexity term whose minimum value is obtained when $\forall j \in \{1, 2\}, \alpha_j = \beta_j$. This can be achieved by assigning equal weight to each sample point in the multi-source data. Based on the analysis above, we can minimize the target error $\epsilon_T(\hat{h})$ by MTM.

Conclusion

In this work, we deeply investigate UDAOD methods based on DETR and uncover several issues. Previous one-stage feature alignment methods overlook their inherent issues: performance fluctuation and training stagnation, while the two-stage feature alignment method based on mean teacher introduces new challenges like unreliable pretrained model and unstable performance gains. Besides, how to utilize the third related domain such as the target-like domain to assist domain adaptation remains unexplored in existing methods. To address the issues and build a robust domain adaptive Detection Transformer framework, we propose a two-stage framework named MTM, *i.e.* Mean Teacher DETR with Masked Feature Alignment.

In the pretraining stage, CycleGAN is used to generate target-like images. We incorporate generated target-like images in pretraining to avoid performance fluctuation. In the self-training stage, we leverage unlabeled target images by pseudo labels based on mean teacher. We propose Object Queries Knowledge Transfer (OQKT) to achieve consistent performance gains. Above all, we propose masked feature alignment including Masked Domain Query-based Feature Alignment (MDQFA) and Masked Token-Wise Feature Alignment (MTWFA) to alleviate domain shift in a more robust way. Our masked feature alignment methods not only prevent training stagnation and lead to a robust pretrained model in the pretraining stage but also enhance the model's final target performance in the self-training stage.

Experimental results and a theoretical analysis have proven the effectiveness of MTM. We expect that our research will inspire future work in this area.

Acknowledgements

References

Anthony, M.; Bartlett, P. L.; Bartlett, P. L.; et al. 1999. *Neural network learning: Theoretical foundations*, volume 9. cambridge university press Cambridge.

Arjovsky, M.; Chintala, S.; and Bottou, L. 2017. Wasserstein generative adversarial networks. In *International conference on machine learning*, 214–223. PMLR.

Arruda, V. F.; Paixao, T. M.; Berriel, R. F.; De Souza, A. F.; Badue, C.; Sebe, N.; and Oliveira-Santos, T. 2019. Cross-domain car detection using unsupervised image-to-image translation: From day to night. In *2019 International Joint Conference on Neural Networks (IJCNN)*, 1–8. IEEE.

Bartlett, P. L.; Foster, D. J.; and Telgarsky, M. J. 2017. Spectrally-normalized margin bounds for neural networks. *Advances in neural information processing systems*, 30.

Ben-David, S.; Blitzer, J.; Crammer, K.; Kulesza, A.; Pereira, F.; and Vaughan, J. W. 2010. A theory of learning from different domains. *Machine learning*, 79: 151–175.

Blitzer, J.; Crammer, K.; Kulesza, A.; Pereira, F.; and Wortman, J. 2007. Learning bounds for domain adaptation. *Advances in neural information processing systems*, 20.

Bousmalis, K.; Trigeorgis, G.; Silberman, N.; Krishnan, D.; and Erhan, D. 2016. Domain separation networks. *Advances in neural information processing systems*, 29.

Cai, Q.; Pan, Y.; Ngo, C.-W.; Tian, X.; Duan, L.; and Yao, T. 2019. Exploring object relation in mean teacher for cross-domain detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 11457–11466.

Carion, N.; Massa, F.; Synnaeve, G.; Usunier, N.; Kirillov, A.; and Zagoruyko, S. 2020. End-to-end object detection with transformers. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part I 16*, 213–229. Springer.

Chen, C.; Zheng, Z.; Ding, X.; Huang, Y.; and Dou, Q. 2020. Harmonizing transferability and discriminability for adapting object detectors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8869–8878.

Chen, M.; Chen, W.; Yang, S.; Song, J.; Wang, X.; Zhang, L.; Yan, Y.; Qi, D.; Zhuang, Y.; Xie, D.; and Pu, S. 2022. Learning Domain Adaptive Object Detection with Probabilistic Teacher. arXiv:2206.06293.

Chen, Y.; Li, W.; Sakaridis, C.; Dai, D.; and Van Gool, L. 2018. Domain adaptive faster r-cnn for object detection in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 3339–3348.

Cordts, M.; Omran, M.; Ramos, S.; Rehfeld, T.; Enzweiler, M.; Benenson, R.; Franke, U.; Roth, S.; and Schiele, B. 2016. The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 3213–3223.

Crammer, K.; Kearns, M.; and Wortman, J. 2008. Learning from Multiple Sources. *Journal of Machine Learning Research*, 9(8).

Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, 248–255. Ieee.

Deng, J.; Li, W.; Chen, Y.; and Duan, L. 2021. Unbiased mean teacher for cross-domain object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4091–4101.

Ding, L.; Wang, L.; and Tao, D. 2020. Self-Attention with Cross-Lingual Position Representation. arXiv:2004.13310.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; Uszkoreit, J.; and Houlsby, N. 2021. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

Ganin, Y.; Ustinova, E.; Ajakan, H.; Germain, P.; Larochelle, H.; Laviolette, F.; Marchand, M.; and Lempitsky, V. 2016. Domain-adversarial training of neural networks. *The journal of machine learning research*, 17(1): 2096–2030.

Girshick, R. 2015. Fast r-cnn. In *Proceedings of the IEEE international conference on computer vision*, 1440–1448.

Gong, K.; Li, S.; Li, S.; Zhang, R.; Liu, C. H.; and Chen, Q. 2022. Improving Transferability for Domain Adaptive Detection Transformers. In *Proceedings of the 30th ACM International Conference on Multimedia*, 1543–1551.

Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; and Bengio, Y. 2020. Generative adversarial networks. *Communications of the ACM*, 63(11): 139–144.

Guo, T.; Huynh, C. P.; and Solh, M. 2019. Domain-adaptive pedestrian detection in thermal images. In *2019 IEEE international conference on image processing (ICIP)*, 1660–1664. IEEE.

He, F.; Liu, T.; and Tao, D. 2020. Why resnet works? residuals generalize. *IEEE transactions on neural networks and learning systems*, 31(12): 5349–5362.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.

He, M.; Wang, Y.; Wu, J.; Wang, Y.; Li, H.; Li, B.; Gan, W.; Wu, W.; and Qiao, Y. 2022. Cross domain object detection by target-perceived dual branch distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9570–9580.

Johnson-Roberson, M.; Barto, C.; Mehta, R.; Sridhar, S. N.; Rosaen, K.; and Vasudevan, R. 2017. Driving in the Matrix: Can Virtual Worlds Replace Human-Generated Annotations for Real World Tasks? arXiv:1610.01983.

Krizhevsky, A.; Sutskever, I.; and Hinton, G. E. 2017. Imagenet classification with deep convolutional neural networks. *Communications of the ACM*, 60(6): 84–90.

Li, W.; Li, F.; Luo, Y.; Wang, P.; et al. 2020. Deep domain adaptive object detection: a survey. In *2020 IEEE Symposium Series on Computational Intelligence (SSCI)*, 1808–1813. IEEE.

Liu, S.; John, V.; Blasch, E.; Liu, Z.; and Huang, Y. 2018. IR2VI: Enhanced night environmental perception by unsupervised thermal image translation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 1153–1160.

Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S.; Fu, C.-Y.; and Berg, A. C. 2016. Ssd: Single shot multibox detector. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14*, 21–37. Springer.

Lv, W.; Zhao, Y.; Xu, S.; Wei, J.; Wang, G.; Cui, C.; Du, Y.; Dang, Q.; and Liu, Y. 2023. DETRs Beat YOLOs on Real-time Object Detection. arXiv:2304.08069.

Menke, M.; Wenzel, T.; and Schwung, A. 2022. AWADA: Attention-Weighted Adversarial Domain Adaptation for Object Detection. arXiv:2208.14662.

Pan, S. J.; Tsang, I. W.; Kwok, J. T.; and Yang, Q. 2010. Domain adaptation via transfer component analysis. *IEEE transactions on neural networks*, 22(2): 199–210.

Redmon, J.; Divvala, S.; Girshick, R.; and Farhadi, A. 2016. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 779–788.

Ren, S.; He, K.; Girshick, R.; and Sun, J. 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28.

Saito, K.; Ushiku, Y.; Harada, T.; and Saenko, K. 2019. Strong-weak distribution alignment for adaptive object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6956–6965.

Sakaridis, C.; Dai, D.; and Van Gool, L. 2018. Semantic foggy scene understanding with synthetic data. *International Journal of Computer Vision*, 126: 973–992.

Shen, Z.; Maheshwari, H.; Yao, W.; and Savvides, M. 2019. Scl: Towards accurate domain adaptive object detection via gradient detach based stacked complementary losses. *arXiv preprint arXiv:1911.02559*.

Simonyan, K.; and Zisserman, A. 2015. Very Deep Convolutional Networks for Large-Scale Image Recognition. In Bengio, Y.; and LeCun, Y., eds., *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Sohn, K.; Zhang, Z.; Li, C.-L.; Zhang, H.; Lee, C.-Y.; and Pfister, T. 2020. A Simple Semi-Supervised Learning Framework for Object Detection. arXiv:2005.04757.

Tzeng, E.; Hoffman, J.; Saenko, K.; and Darrell, T. 2017. Adversarial discriminative domain adaptation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 7167–7176.

Van der Maaten, L.; and Hinton, G. 2008. Visualizing data using t-SNE. *Journal of machine learning research*, 9(11).

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Wang, W.; Cao, Y.; Zhang, J.; He, F.; Zha, Z.-J.; Wen, Y.; and Tao, D. 2021. Exploring sequence feature alignment for domain adaptive detection transformers. In *Proceedings of the 29th ACM International Conference on Multimedia*, 1730–1738.

Xu, C.-D.; Zhao, X.-R.; Jin, X.; and Wei, X.-S. 2020a. Exploring categorical regularization for domain adaptive object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 11724–11733.

Xu, M.; Wang, H.; Ni, B.; Tian, Q.; and Zhang, W. 2020b. Cross-domain detection via graph-induced prototype alignment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 12355–12364.

Xu, M.; Zhang, Z.; Hu, H.; Wang, J.; Wang, L.; Wei, F.; Bai, X.; and Liu, Z. 2021. End-to-end semi-supervised object detection with soft teacher. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 3060–3069.

Yu, F.; Chen, H.; Wang, X.; Xian, W.; Chen, Y.; Liu, F.; Madhavan, V.; and Darrell, T. 2020. BDD100K: A Diverse Driving Dataset for Heterogeneous Multitask Learning. arXiv:1805.04687.

Yu, J.; Liu, J.; Wei, X.; Zhou, H.; Nakata, Y.; Gudovskiy, D. A.; Okuno, T.; Li, J.; Keutzer, K.; and Zhang, S. 2022. MTTrans: Cross-domain Object Detection with Mean Teacher Transformer. In Avidan, S.; Brostow, G. J.; Cissé, M.; Farinella, G. M.; and Hassner, T., eds., *Computer Vision - ECCV 2022 - 17th European Conference, Tel Aviv, Israel, October 23-27, 2022, Proceedings, Part IX*, volume 13669 of *Lecture Notes in Computer Science*, 629–645. Springer.

Zhang, J.; Huang, J.; Luo, Z.; Zhang, G.; Zhang, X.; and Lu, S. 2023. DA-DETR: Domain Adaptive Detection Transformer With Information Fusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 23787–23798.

Zhao, L.; and Wang, L. 2022. Task-specific inconsistency alignment for domain adaptive object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 14217–14226.

Zhu, J.-Y.; Park, T.; Isola, P.; and Efros, A. A. 2017. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, 2223–2232.

Zhu, X.; Su, W.; Lu, L.; Li, B.; Wang, X.; and Dai, J. 2021. Deformable DETR: Deformable Transformers for End-to-End Object Detection. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.