Masked Retraining Teacher-Student Framework for Domain Adaptive Object Detection

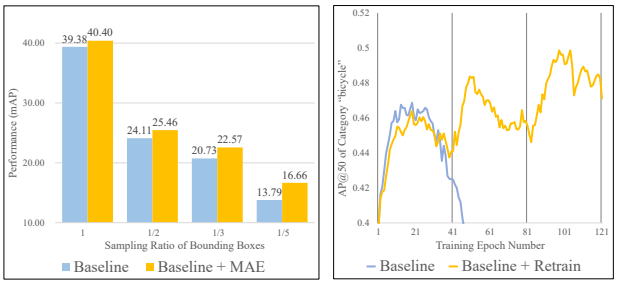
Zijing Zhao[1] Sitong Wei[1] Qingchao Chen[2] Dehui Li[3] Yifan Yang[3] Yuxin Peng[1] Yang Liu[1*]
[1]Wangxuan Institute of Computer Technology, Peking University
[2]National Institute of Health Data Science, Peking University [3]Tencent Intelligent Mobility
zijingzhao@stu.pku.edu.cn {weisitong, qingchao.chen, pengyuxin, yangliu}@pku.edu.cn
{dehuili, lvanyang}@tencent.com

Abstract

Domain adaptive Object Detection (DAOD) leverages a labeled domain (source) to learn an object detector generalizing to a novel domain without annotation (target). Recent advances use a teacher-student framework, i.e., a student model is supervised by the pseudo labels from a teacher model. Though great success, they suffer from the ***limited*** *number of pseudo boxes with* ***incorrect*** *predictions caused by the domain shift, misleading the student model to get sub-optimal results. To mitigate this problem, we propose Masked Retraining Teacher-student framework (MRT) which leverages masked autoencoder and selective retraining mechanism on detection transformer. Specifically, we present a customized design of masked autoencoder branch, masking the multi-scale feature maps of target images and reconstructing features by the encoder of the student model and an auxiliary decoder. This helps the student model capture target domain characteristics and become a more data-efficient learner to gain knowledge from the* ***limited*** *number of pseudo boxes. Furthermore, we adopt selective retraining mechanism, periodically re-initializing certain parts of the student parameters with masked autoencoder refined weights to allow the model to jump out of the local optimum biased to the* ***incorrect*** *pseudo labels. Experimental results on three DAOD benchmarks demonstrate the effectiveness of our method. Code can be found at https://github.com/JeremyZhao1998/MRT-release.*

1. Introduction

Object detection has a wide range of real-world application scenarios, and has been deeply studied in computer vision researches. CNN-based [35, 32, 40] and transformer-based [3, 51] detectors have shown great success in challenging benchmarks. However, they suffer from domain shift where there is an obvious distribution gap between the pretraining data and the deployed environment.



(a) mAP under limited boxes (b) Category AP during training

Figure 1. **(a) Performance under limited number of box annotations**. We adopt MAE in training target images with ground truth labels, but manually reduce the amount of bounding boxes. MAE boosts the performance by a large margin under limited boxes. **(b) Performance of Category "bicycle" during training**. The performance declines rapidly due to the local optimum caused by incorrect pseudo boxes. With the help of retraining (every 40 epochs), the model is able to jump out of the local optimum.

To mitigate the performance drop caused by domain shift without extra annotation, unsupervised domain adaptation (UDA) has been studied in classification, segmentation and object detection tasks, where the model is trained on a labeled *source* domain and an unlabeled *target* domain, and is expected to generalize well on the target domain. As a branch of UDA, unsupervised domain adaptive object detection (DAOD) researches [9, 49, 29, 48] utilize numerous techniques such as adversarial alignment, image-to-image translation, GNNs and mean teacher training, widely improving domain adaptation performance of object detectors.

Among these approaches, [29, 48, 5, 20] use a teacher-student framework where a teacher model produces pseudo labels of the unlabeled target images to supervise a student model, and has achieved significant performance gains.

*Corresponding author

However, the teaching process faces the challenge of low-quality pseudo labels in which there are limited number of pseudo boxes and incorrect predictions caused by the domain shift. Under such framework, pseudo labels are selected from outputs of the teacher model by a threshold of confidence scores. Selecting large amounts of pseudo labels with lower threshold brings too many incorrect predictions, degrading the performance, while with higher threshold, the limited number of pseudo boxes provides sub-optimal supervision. Blue bars of Figure 1(a) shows that even supervised by ground truth labels without domain shift, the performance drops significantly when the amount of box annotations are reduced. Anyhow, incorrect predictions always exist, leading the model to get stuck at the local optimum. As is shown in Figure 1(b), in the later stage of training, performance of some categories declines rapidly (the blue curve) due to the growing number of incorrect pseudo boxes. Though [29] utilizes adversarial alignment and weak-strong augmentation to minimize the false positive ratio of pseudo labels, they ignore the sub-optimal supervision of the limited number of pseudo boxes, and the impact of incorrect pseudo labels which always exist.

To address this issue, we propose Masked Retraining Teacher-student framework (MRT) which is built on the baseline of adaptivee teacher-student framework for Deformable DETR[51] detector, leveraging masked autoencoder (MAE) and selective retraining mechanism.

Unlike pretraining MAEs[19, 41] which leverage large-scale training data under a pretrain-finetune paradigm, we present a customized design of MAE branch, randomly masking portions of multi-scale feature maps of the target images and reconstructing the missing features from their contexts by the encoder of the student model and an auxiliary decoder, simultaneously with the detection loss. As a self-supervised task on target images, MAE leads the transformer encoder to gain more intimate knowledge of the target domain from the limited number of pseudo boxes. Empirically, we observe that MAE helps the model encode better features, improving the performance under any amount of supervision and achieving larger gains when fewer box annotations are provided, as is shown in Figure 1(a).

Furthermore, as the student model is sensitive to the pseudo label noise, we adopt a simple yet effective selective retraining mechanism. Specifically, we periodically re-initialize certain parts of the student parameters to allow the model to jump out of the local optimum biased to the incorrect pseudo labels. Unlike existing retraining approaches [18, 34] that ignore the quality of re-initialization weights, we re-initialize the student model with MAE refined weights to avoid low-quality student weights to impact the teacher model through EMA. In teacher-student framework, an enhanced teacher model helps the student recover rapidly after the re-initialization. As is shown in the orange curve of Figure 1(b), retraining recovers the performance that has been corrupted by noisy pseudo labels.

We summarize the contribution of this paper as follow: 1) We propose a novel Masked Retraining Teacher-student framework(MRT) for training domain adaptive detection transformers, which is built on the adaptive teacher-student framework baseline and overcomes the problem of low-quality pseudo labels. 2) To the best of our knowledge, we are the first to present that masked autoencoder is a data-efficient domain adapter which helps the model better capture domain characteristics, and the first to adopt selective retraining mechanism in teacher-student framework to help the model jump out of local optimums. 3) Our method outperforms existing approaches by a large margin and achieves state-of-the-art on three DAOD benchmarks.

2. Related Work

Object detection: Object detection has been deeply studied in computer vision. CNN-based methods[35, 30, 31] present region proposals networks and achieve outstanding performance. One-stage detectors without region proposals [32, 40] simplify the structure and perform faster. Recently, transformer-based models [3, 51] are also developed in object detection, exploring token-wise dependencies for context modeling. These methods all focus on supervised learning, while we aim to generalize the model to a novel domain without extra annotations. In this work, we employ Deformable DETR[51] as the detector due to its simplified one-stage structure and the flexible transfer-learning ability of the transformer architecture.

Domain adaptive object detection: Domain adaptive object detection (DAOD) is raised to overcome the domain shift problem in object detection. As a pioneer work of DAOD, [9] investigates adversarial feature alignment for Faster R-CNN. Following [9], [21, 8, 27, 7, 4, 36, 46] apply different aspects of feature alignments, and [44, 24, 17] elaborately design alignments for transformer architectures. Numerous techniques such as image-to-image translation[49], graph reasoning[28, 48] and pseudo label self-training[12, 29, 48] have also been studied for DAOD. Among existing approaches, methods utilizing teacher-student framework achieve leading position in experimental performances. [29] utilizes adversarial alignment and weak-strong augmentation to minimize the false positive ratio of pseudo labels. [20] reduces domain shift by incorporating target object knowledge through self-distillation. [5] employs uncertainty-guided self-training to promote both classification and localization adaptations. Though great success, they ignore the low-quality pseudo labels which contains limited number of bounding boxes and incorrect predictions. We build our method on the baseline of teacher-student framework with adversarial alignment, introducing MAE and selective retraining to overcome such problem.

Masked autoencoders: Autoencoding is a classical representation learning method, utilizing an encoder to map the input to a latent representation and a decoder to reconstruct the input. Denoising autoencoders (DAE)[42] corrupt the input signal and learn to reconstruct the original, uncorrupted signal. Recently, DAE methods are widely used in large-scale pretraining tasks in both language models [13, 2] and vision models [6, 14, 1], holding out a portion of the input and train models to predict the missing content. The influential MAE[19] randomly masks image patches with a high portion and reconstruct the missing pixels by an asymmetric architecture to pretrain a vision transformer that generalize well in downstream tasks. [41] explores MAE in video pretraining, raising the point that MAE is a data-efficient learner and domain shift is an important factor during pretraining. [11] employs attention mask for DETR decoder as an unsupervised pretraining task. However, these works all focus on pretraining transformer-based backbones with all accessible data to generalize to downstream tasks, without considering domain shift. We empirically observe that feeding MAE with inputs from mismatched domain provides sub-optimal results. Thus, we apply MAE only on target images for domain adaptation, guiding the encoder to better capture target domain characteristics.

Overcoming local optimums: Optimizing neural networks faces the challenge of local optimums. Regularization methods such as Dropout[38] and DropBlock[16] try to avoid the model from over-fitting or getting trapped in local minimums by introducing randomness. [23] introduces cyclic learning rate, but is inapplicable for transformer-based models. Recent studies show that retraining is a simple yet effective way for transformers. [18] proposes DSD retraining with reference to model pruning to avoid over-fitting to noisy data, and [34] proposes selective retraining mechanism for visual grounding transformers to converge to better minimums. To the best of our knowledge, we are the first to adopt retraining in teacher-student framework to overcome the impact of incorrect pseudo labels. Under such framework, quality of the re-initialization weights become significant, which has not been studied in existing retraining approaches, since the teacher model may be influenced by the re-initialized student weights to provide sub-optimal pseudo labels. We re-initializing the student parameters with MAE refined weights to avoid the problem.

3. Problem Formulation and Baseline

3.1. Problem Formulation

We first review the problem formulation of unsupervised DAOD. Given a labeled source dataset $D_s = \{(x_s^i, y_s^i)\}_{i=1}^{N_s}$ sized N_s and an unlabeled target dataset $D_t = \{x_t^i\}_{i=1}^{N_t}$ sized N_t, where x denotes an image and $y = (b, c)$ represents an annotation for object detection including bounding box b and the corresponding category c, we train a domain adaptive detector with both D_s and D_t, and evaluate the detection performance on data in the target domain.

3.2. Adaptive Teacher-student Baseline Revisited

Recent advances [46, 48] use a teacher-student framework combined with adversarial alignment and achieve SOTA. Our method is built on the adaptive teacher-student framework as baseline. We revisit the baseline as follow.

Teacher-student framework: A teacher model T and a student model S share the same structure of backbone, encoder and decoder. The teacher takes a weakly augmented target image x_t and produces pseudo labels $(\hat{b_t}, \hat{c_t})$. The student takes both source and target images which are strongly augmented. Supervised loss $\mathcal{L}_{sup}$ is calculated on x_s with their ground truth labels same as [51]:

$$\mathcal{L}_{sup} = \mathcal{L}_{box}^{S}(x_s, y_s) + \mathcal{L}_{giou}^{S}(x_s, y_s) + \mathcal{L}_{cls}^{S}(x_s, y_s) \quad (1)$$

while x_t receive supervision from pseudo labels, but only in classification task following [46] as unsupervised loss:

$$\mathcal{L}_{unsup} = \mathcal{L}_{cls}^{S}(x_t, \hat{b_t}, \hat{c_t}) \quad (2)$$

The teacher is only updated by Exponential Moving Average (EMA) from the student without gradient accumulation:

$$\theta_t \leftarrow \alpha\theta_t + (1 - \alpha)\theta_s \quad (3)$$

where θ_t and θ_s denotes the model parameters of teacher and student respectively, and α is a hyper-parameter.

Discriminators for adversarial alignment: Domain discriminators D are placed after certain components to predict the domain label of the features, updated by BCE loss. On Deformable DETR detector, we set discriminators for the backbone, encoder and decoder to provide global level, token-wise multi-scale feature level and instance level alignment respectively, and $\mathcal{L}_{dis}$ denotes their weighted sum. The adversarial optimization objective is formulated:

$$\mathcal{L}_{adv} = \max_{S} \min_{D} \mathcal{L}_{dis} \quad (4)$$

where S, D denotes the student and discriminators respectively. Gradient Reverse Layers(GRL)[15] is adopted for min-max optimization. Overall objective of the student is:

$$\mathcal{L}_{teach} = \mathcal{L}_{sup} + \mathcal{L}_{adv} + \lambda_{unsup}\mathcal{L}_{unsup} \quad (5)$$

where λ_{unsup} is the hyper-parameter, while the teacher is updated only by EMA which has been discussed. Before teaching process, the model will first be trained with $\mathcal{L}_{sup}$ using annotated source data D_s, and both teacher and student model will be initialized with such parameters.

The adaptive teacher-student baseline on Deformable DETR outperforms most existing DAOD approaches. However, the problem of low-quality pseudo labels still exist,

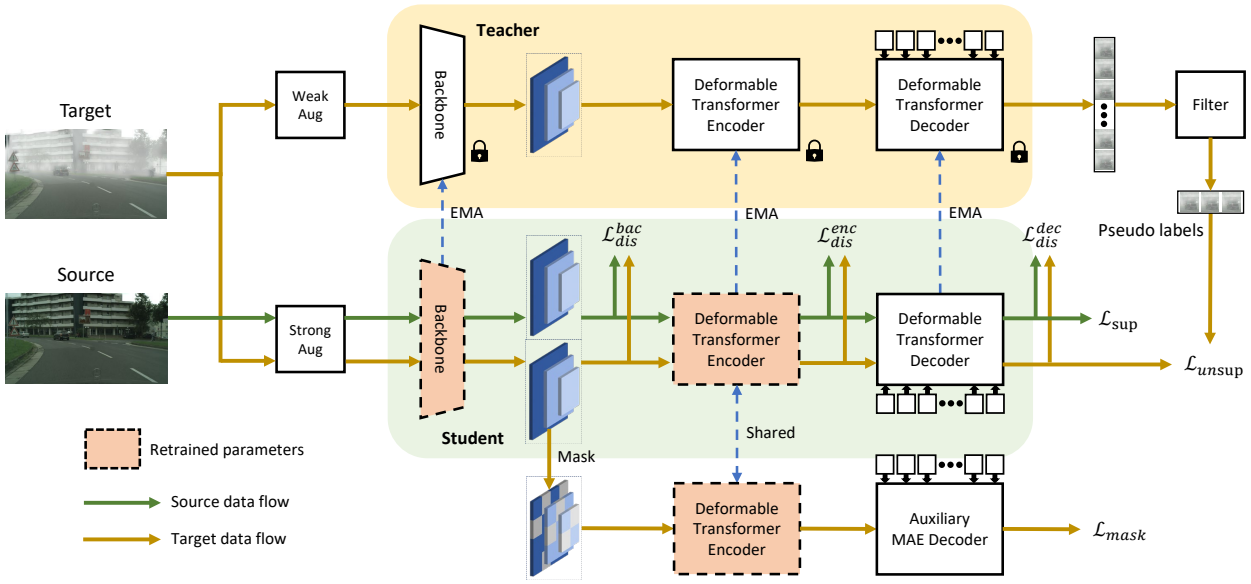


Figure 2. **Overview Masked Retraining Teacher-student framework(MRT).** The adaptive teacher-student baseline consists of a teacher model which takes weakly-augmented target images and produces pseudo labels, and a student model which takes strongly augmented source and target images, supervised by ground truth labels and pseudo labels respectively. Adversarial alignment are applied on backbone, encoder and decoder. Our proposed MAE branch masks feature maps of target images, and and reconstructs the feature by student encoder and an auxiliary decoder. Our proposed selective retraining mechanism periodically re-initialize certain parts of the student parameters as highlighted. The teacher model is updated only by EMA from the student model. Empirically, we use the teacher model at inference time.

leading the model to sub-optimal results, as has been shown in Figure 1. Our proposed method overcomes this issue by introducing MAE branch and selective retraining mechanism which will be elaborated in the following section.

4. Proposed Method

4.1. Method Overview

Our proposed method can be overviewed in Figure 2, which is built on the adaptive teacher-student baseline discussed in Section 3.2. We propose a customized design of **masked autoencoder(MAE) branch** for student model, randomly masking a portion of the multi-scale feature maps of the target images, feeding them into the encoder, and introducing an auxiliary decoder to reconstruct the missing features from their contexts. As a self-supervised task on target images, MAE leads the transformer encoder to gain more intimate knowledge of the target domain, thus helps the model better capture target domain characteristics and boosts the performance when the amount of pseudo boxes are limited. Employing **selective retraining mechanism**, we periodically re-initialize parts of the student parameters with the MAE refined weights, as is illustrated in Figure 2 with marked units, while the teacher model is continually updated via exponential moving average (EMA) from the student model. In this way, the teacher model continually produces relatively high-quality pseudo labels without being corrupted by re-initialized low-quality student weights, and the student model is allowed to jump out of the local optimum biased to the incorrect pseudo boxes. Selecting pseudo labels from outputs of the teacher is also important. A fixed confidence threshold is often used, but is unaware of the imbalance sample class distribution and the training period. We propose a **dynamic threshold** strategy to set different thresholds for each category dynamically during training, selecting appropriate amount of pseudo boxes.

4.2. Masked Autoencoder Branch

We propose a customized design of MAE branch for the student model to better capture target domain characteristic as a self-supervised domain adapter, and to effectively gain knowledge from the limited number of high-quality pseudo boxes. The multi-scale feature maps of target images will be masked and fed into the student encoder, and reconstructed by an auxiliary decoder. The masking and reconstructing processes are elaborated as follows.

Feature masking: Unlike ViT [14] backbone which directly takes image patches as the input, the Deformable DETR encoder takes the multi-scale feature maps $\{z_i \in \mathbb{R}^{C_i \times H_i \times W_i}\}_{i=1}^{K}$ which are projected from the output of the backbone, where K denotes the number of scales. We randomly generate masking $\{m_i \in \{0,1\}^{H_i \times W_i}\}_{i=1}^{K}$ for every scale of the feature maps with the same masking ratio. The deformable attention rely on the spatial structure of the

feature maps to generate reference points and sampling locations that is used to choose points on the value matrix to perform sparse attention, so the spatial structure of the input cannot be serialized. As a result, we set the masked portion of the corresponding value matrix to zero in deformable attention to perform zero-masking, unlike [19] that only sends unmasked patches as a serialized input to ViT.

Reconstruction: The encoder takes the masked feature maps, and further encodes the feature maps with deformable attention. On the output features of the encoder, we fill the masked portion with a shared mask query q_m and send them to an auxiliary MAE decoder S_m to reconstruct the feature map. Following [19], we adopt an asymmetric lightweight decoder. Since the last layer of the backbone outputs, i.e., z_K contains all the semantic information, we only reconstruct it to speed up convergence and reduce computation. The last layer of MAE decoder is a linear projection whose output channels equals to the channels of z_K. Eventually, we compute mean square error (MSE) between the reconstructed and original features, but only on the masked portion, formulated as:

$$\hat{z_K} = S_m(S_e(mask(m, x_t)), q_m) \quad (6)$$

$$\mathcal{L}_{mask} = \mathcal{L}_{MSE}(mask(m_K, \hat{z_K}), mask(m_K, z_K)) \quad (7)$$

where S_e, S_m, q_m denotes the encoder of student model, the auxiliary MAE decoder and the mask query respectively.

Training data: Note that although student model takes both source and target images, we employ MAE only on target images, since DAOD aims to train the detector to perform better on the unlabeled target domain, and applying MAE with source images leads the model to generate source domain features with sub-optimal performance. Employing MAE on target images helps the student encoder to gain more intimate knowledge of the target domain, and boosts performance when high-quality pseudo labels are limited. We apply MAE on student model instead of teacher, for it is the student model who suffer from sub-optimal supervision of the limited pseudo boxes. Besides, the teacher model is not supervised by any detection loss. Employing MAE on the teacher model creates an encoder that is incompatible with the decoder and detection head, thus is unable to provide meaningful pseudo labels.

Training strategy: We train the MAE branch and detection loss simultaneously instead of following the pretrain-finetune paradigm, since pretraining leads the model to over-fit to reconstruction task due to the relatively small dataset with distinctive domain characteristics. Before the teaching process, we first train the model with $\mathcal{L}_{sup}$ in Equation(1) together with $\mathcal{L}_{mask}$ in Equation(7) to produce an enhanced initialization θ_{mask} for teaching, and as a refined re-initialization for selective retraining. The overall objective of student model in teaching process is:

$$\mathcal{L} = \mathcal{L}_{teach} + \lambda_{mask}\mathcal{L}_{mask} \quad (8)$$

where $\mathcal{L}_{teach}$ has been formulated in Equation(5), and λ_{mask} is the coefficient of $\mathcal{L}_{mask}$ in Equation(7). As the teaching process continues, adequate amount of pseudo boxes will be produced, and the influence of $\mathcal{L}_{mask}$ should be lowered to avoid the encoder over-fitting to reconstruction task. As a result, We decay λ_{mask} as teaching continues. More details will be discussed in Section 5.4.

4.3. Selective Retraining Mechanism

The transformer-based models tend to over-fit without large-scale training data, especially when noisy annotations are included. Here in our case, pseudo labels always contain noise since they are provided by the updating teacher model instead of ground truths. As is shown in Figure 1(b), in later stage of the teaching process, performance of certain categories are severely influenced due to the over-fitting to the incorrect pseudo labels. Worse still, since the teacher model is continually updated via EMA from student model, it could also by affected and produces even worse pseudo labels. We adopt selective retraining mechanism on student branch to help student model jump out of local optimums biased to incorrect pseudo labels.

To be specific, during teaching period, we continually update the teacher model via EMA without retraining it, but periodically re-initialize parts of the student model with the MAE refined parameters θ_{mask} which has been discussed in Section 4.2. Though θ_{mask} do not contain detection knowledge, its target domain encoding ability has not been corrupted by noisy pseudo labels. After re-initialization, the retrained components are allowed to leave the local optimum and can be guided by the fixed parts which can be regarded as better trained, and by an enhanced teacher which is continually updated. We do not retrain the teacher model since re-initialization cause a sudden drop of pseudo label quality, thus deteriorate the teaching process. We re-initialize the student model with MAE refined weights θ_{mask} rather than source-only trained weights or random weights for a quicker recovery, preventing its negative influence to teacher via EMA. We choose the retrained components by experimental studies. We try to perform re-initialization on different parts of the student model and choose the setting with best performance in which the decoder is kept while the backbone and encoder are retrained. More details will be discussed in Section 5.4.

4.4. Dynamic Threshold

Existing teacher-student frameworks on DAOD set a fixed confidence threshold to select pseudo labels from outputs of the teacher model. However, we observe in experiments that as the teaching process continues, the predicted confidence scores tend to rise, introducing too many pseudo labels which contain much more incorrect labels. Besides, fixed threshold for every category ignores the category dis-

Method	Detector	person	rider	car	truck	bus	train	mcycle	bicycle	mAP
FasterRCNN[35](Source)	FRCNN	26.9	38.2	35.6	18.3	32.4	9.6	25.8	28.6	26.9
DA-Faster[9]	FRCNN	29.2	40.4	43.4	19.7	38.3	28.5	23.7	32.7	32.0
UMT[12]	FRCNN	33.0	46.7	48.6	34.1	56.5	46.8	30.4	37.3	41.7
TIA[50]	FRCNN	34.8	46.3	49.7	31.1	52.1	48.6	37.7	38.1	42.3
D-adapt[25]	FRCNN	40.8	47.1	57.5	33.5	46.9	41.4	33.6	43.0	43.0
SIGMA[28]	FRCNN	44.0	43.9	60.3	31.6	50.4	51.5	31.7	40.6	44.2
AT[1][29]	FRCNN	43.7	**54.1**	62.3	31.9	54.4	49.3	35.2	47.9	47.4
TDD[20]	FRCNN	50.7	53.7	68.2	35.1	53.0	45.1	38.9	**49.1**	49.2
PT[5]	FRCNN	40.2	48.8	63.4	30.7	51.8	30.6	35.4	44.5	42.7
FCOS[40] (Source)	FCOS	36.9	36.3	44.1	18.6	29.3	8.4	20.3	31.9	28.2
EPM[22]	FCOS	44.2	46.6	58.5	24.8	45.2	29.1	28.6	34.6	39.0
SSAL[33]	FCOS	45.1	47.4	59.4	24.5	50.0	25.7	26.0	38.7	39.6
Def DETR[51] (Source)	Def DETR	37.7	39.1	44.2	17.2	26.8	5.8	21.6	35.5	28.5
SFA[44]	Def DETR	46.5	48.6	62.6	25.1	46.2	29.4	28.3	44.0	41.3
MTTrans[48]	Def DETR	47.7	49.9	65.2	25.8	45.9	33.8	32.6	46.5	43.4
O^2net[17]	Def DETR	48.7	51.5	63.6	31.1	47.6	47.8	38.0	45.9	46.8
AQT[24]	Def DETR	49.3	52.3	64.4	27.7	53.7	46.5	36.0	46.4	47.1
MRT(Ours)	Def DETR	**52.8**	51.7	**68.7**	**35.9**	**58.1**	**54.5**	**41.0**	47.1	**51.2**

Table 1. **Results of *Cityscapes* to *Foggy Cityscapes(0.02)*.** "FRCNN" denotes Faster R-CNN and "Def DETR" denotes Deformable DETR.

tribution of samples. To address this issue, we initialize the thresholds for each category by a same value, and dynamically update them based on the predicted confidence scores of the source domain instances. In this way, the thresholds for each category could be updated differently. More details and discussion of the rationality and effectiveness of this method can be found in Appendix.

5. Experiments

5.1. Datasets

Following existing DAOD approaches [44, 48, 17, 24], we evaluate our method on the following benchmarks.

Cityscapes to Foggy Cityscapes: Cityscapes[10] is collected from urban scenes containing 2,975 images for training and 500 images for validation. Foggy Cityscapes[37] is constructed by a fog synthesis algorithm from Cityscapes. We use Cityscapes as source domain and Foggy Cityscapes with highest fog density(0.02) as target domain.

Cityscapes to BDD100k-daytime: BDD100k[47] is a large-scale driving dataset. Its daytime subset containing 36,728 training images and 5,258 validation images is commonly used in DAOD. We used Citycapes as source domain and BDD100k-daytime as target domain.

Sim10k to Cityscapes(car): Sim10k[26] is a synthetic dataset from GTA game engine containing 10,000 images. We use Sim10k as source domain and "car" instances in Cityscapes as target domain.

[1]We run *Foggy(0.02)* on AT's open source code to get the results instead of its originally reported *Foggy(all)* for a fair comparison.

5.2. Implementation Details

We use Deformable DETR [51] as our base detector. For loss coefficients, we set $\lambda_{unsup} = 1.0$ and initial $\lambda_{mask} = 1.0$. For the discriminators, the coefficients of backbone, encoder and decoder is set $0.3, 1.0, 1.0$ respectively. We set the weight smooth parameter $\alpha = 0.9996$ in EMA. For dynamic threshold, we initialize the thresholds for each category by 0.3. For MAE branch, we use a 2-layer asymmetric decoder and a mask ratio of 0.8. We optimize the network by Adam optimizer with initial learning rate 2×10^{-4} and batch size 8. The data augmentation methods include random horizontal flip for weak augmentation, and randomly color jittering, grayscaling and Gaussian blurring for strong augmentations. Algorithms are implemented by PyTorch. More implementation details which differ between benchmarks can be seen in Appendix.

5.3. Comparing with Other Methods

We compare our proposed MRT with other methods on the three benchmarks mentioned above. Our proposed MRT outperforms previous methods, and achieves significant improvements compared with DETR-based methods. As shown in Table 1, for categories with fewer instances (i.e."truck") in *Cityscapes* to *Foggy Cityscapes*, MRT performs much better, as the data-efficient MAE branch boost the performance. For confusing categories (i.e."bicycle" and "motorcycle"), MRT has a significant performance gain with the help of selective retraining mechanism. As shown in Table 2, on *Cityscapes* to *BDD100k-daytime* where the class distribution of target dataset is largely different from

Method	Detector	person	rider	car	truck	bus	mcycle	bicycle	mAP
FasterRCNN[35](Source)	FRCNN	28.8	25.4	44.1	17.9	16.1	13.9	22.4	24.1
DA-Faster[9]	FRCNN	28.9	27.4	44.2	19.1	18.0	14.2	22.4	24.9
ICR-CCR-SW[45]	FRCNN	32.8	29.3	45.8	22.7	20.6	14.9	**25.5**	27.4
FCOS[40] (Source)	FCOS	38.6	24.8	54.5	17.2	16.3	15.0	18.3	26.4
EPM[22]	FCOS	39.6	26.8	55.8	18.8	19.1	14.5	20.1	27.8
Def DETR[51] (Source)	Def DETR	38.9	26.7	55.2	15.7	19.7	10.8	16.2	26.2
SFA[44]	Def DETR	40.2	27.6	57.5	19.1	23.4	15.4	19.2	28.9
AQT[24]	Def DETR	38.2	**33.0**	58.4	17.3	18.4	16.9	23.5	29.4
O^2net[17]	Def DETR	40.4	31.2	58.6	20.4	25.0	14.9	22.7	30.5
MTTrans[48]	Def DETR	44.1	30.1	61.5	**25.1**	**26.9**	17.7	23.0	32.6
MRT(Ours)	Def DETR	**48.4**	30.9	**63.7**	24.7	25.5	**20.2**	22.6	**33.7**

Table 2. **Results of *Cityscapes* to *BDD100k-daytime*.**

Method	Detector	carAP
Faster R-CNN(Source)[35]	FRCNN	39.4
DA-Faster[9]	FRCNN	41.9
MeGA-CDA[43]	FRCNN	44.8
GPA[45]	FRCNN	47.6
ViSGA[36]	FRCNN	49.3
KTNet[39]	FRCNN	50.7
PT[5]	FRCNN	55.1
FCOS(Source)[40]	FCOS	42.5
EPM[22]	FCOS	47.3
SSAL[33]	FCOS	51.8
Deformable DETR(Source)[51]	Def DETR	47.4
SFA[44]	Def DETR	52.6
AQT[24]	Def DETR	53.4
O^2net[17]	Def DETR	54.1
MTTrans [29]	Def DETR	57.9
MRT(Ours)	Def DETR	**62.0**

Table 3. **Results of *Sim10k* to *Cityscapes(car)*.**

source, MRT still achieves leading. As shown in Table 3, on *Sim10k* to *Cityscapes(car)*, MRT outperforms previous SOTAs, indicating that our proposed method stays effective under larger domain gap.

5.4. Ablation Study and Analysis

In this section, we provide ablation study and analysis of our proposed approaches. All experiments are conducted on *Cityscapes* to *Foggy Cityscapes*.

Quantitative Ablation Study: The effect of proposed modules are presented in Table 4, from which we can observe: 1) Introducing MAE branch without teacher-student framework (line 2) shows a significant improvement compared with source-only trained model, indicating that MAE is a data-efficient domain adapter. 2) MAE branch, selective retraining and dynamic filter respectively improves the performance by a large margin (line 4-6) when independently

Source	Baseline	DT	Retrain	MAE	mAP
✓					28.5
✓				✓	35.8
✓	✓				44.9
✓	✓	✓			45.5
✓	✓	✓		✓	48.3
✓	✓	✓	✓		48.1
✓	✓	✓	✓	✓	**51.2**

Table 4. **Ablation studies of proposed modules on *Cityscapes* to *Foggy Cityscapes*.** "Source" denotes the source-only trained model. "Baseline" denotes the adaptive teacher-student baseline. "DT", "Retrain" and "MAE" denotes proposed dynamic threshold, selective retraining and masked autoencoder branch, respectively.

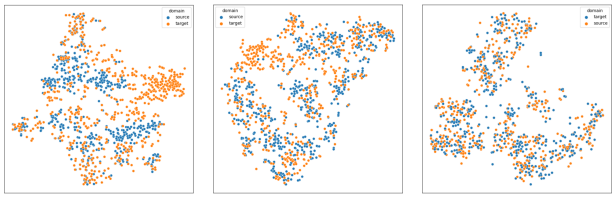
used, illustrating their effectiveness. 3) Retraining combining with MAE refined re-initialization (line 7) further improves the performance, indicating that our proposed design maximizes the advantages of both modules.

Qualitative visualization analysis: We provide qualitative visualization analysis of pseudo labels and feature distributions. Figure 3 shows the selected pseudo labels. From Figure 3(b) to 3(c), MAE guide the model to provide more pseudo labels, and from Figure 3(c) to 3(d), selective retraining filters out incorrect pseudo labels and predicts closer to ground truth. Figure 4 visualizes the features of two domains by t-SNE. MAE branch provides a much better initialization compared to source-only trained model, and our proposed MRT further bridges the domain gap. More visualization figures can be seen in Appendix.

Analysis of MAE branch: Table 5 shows the detailed results of MAE ablation. Table 5(a) indicates that relatively high mask ratio yields a nontrivial and meaningful self-supervisory task. Table 5(b) explores the decay strategy of MAE coefficient λ_{mask}. "no decay" denotes the MAE loss continually exists with fixed λ_{mask}. "linear" denotes λ_{mask} decays linearly during teaching process. "hard" de-

(a) Source-only (b) Baseline (c) MRT w/o Retrain (d) MRT (ours) (e) Ground truth

Figure 3. **Qualitative ablation: pseudo labels for *Foggy Cityscapes*.** "Baseline" denotes the adaptive teacher-student baseline.



(a) Source-only (b) Source+MAE (c) MRT (ours)

Figure 4. **Qualitative ablation: feature visualization of the two domains on *Cityscapes* to *Foggy Cityscapes* by t-SNE.** The blue and orange points denotes source and target features respectively.

Mask ratio	mAP
0.5	47.1
0.8	**48.3**
0.9	46.8

(a) Mask ratio

Strategy	mAP
No decay	47.9
Linear	48.2
Hard	**48.3**

(b) Coefficient decay

Training data	mAP
Source	44.5
Source & Target	47.0
Target	**48.3**

(c) MAE Training data

MAE usage	mAP
w/o MAE	44.9
Pretrain	42.7
MAE branch	**48.3**

(d) MAE usage

Table 5. **Ablation studies of MAE branch.** (a) Results of different mask ratios. (b) Results of coefficient decay strategies. (c) The training data that MAE are applied to. (d) The usage of MAE (as a branch or as an independent pretraining stage).

notes that we drop the MAE branch after a certain teaching epoch. "No decay" receives sub-optimal result for the MAE branch benefits less when more pseudo labels are provided, thus its influence to the overall objective should be lowered at later stage of training. Table 5(c) illustrates that MAE is sensitive to domain shift of the input data. Training with target features receives better performance on target evaluation compared with the inputs that consist source images. Table 5(d) compares the result of MAE branch with the pretrain-finetune paradigm which is commonly used in pretraining tasks, indicating that in DAOD, pretraining on relatively small dataset with distinctive domain characteristics leads the model to over-fit on reconstruction task. MRT use a customised MAE branch to overcome this issue.

Analysis of Selective Retraining: Utilizing selective

Bac.	Enc.	Dec.	mAP
			44.9
✓	✓	✓	46.0
	✓	✓	46.6
✓		✓	47.1
✓	✓		**48.1**

(a) Retrained module

Weights	mAP
random	45.5
source	48.1
source+MAE	**51.2**

(b) Re-initialization weights

Table 6. **Ablation studies of selective retraining.** (a) Results of different retrained modules (tick for retrained). "Bac.", "Enc." and "Dec." denotes backbone, encoder and decoder respectively. (b) Results of different re-initialization weights.

retraining, we periodically re-initialize some components of the model while keeping other components as enhanced parts. In teacher-student framework, we do not re-initialize teacher as has been discussed in Section 4.3. For the student model, we choose the retrained modules through experimental results. Table 6(a) shows that keeping the decoder updated and retraining backbone and decoder gets the best performance. For the re-initialization weights, Table 6(b) shows that MAE refined weights receives the best performance. Further discussion on the selective retraining mechanism can be seen in Appendix.

6. Conclusion

In this paper, we propose a novel Masked Retraining Teacher-student framework (MRT) on domain adaptive object detection task. Our customized design of masked autoencoder branch helps the student model better capture target domain characteristics and gain knowledge from limited amount of pseudo boxes, and the selective retraining mechanism allows the model to jump out of the local optimum biased to the incorrect pseudo labels. Experiments on three benchmarks confirmed the effectiveness of our model. Extensive ablation experiments demonstrate that every design helps to improve domain adaptation ability of the model.

7. Acknowledgements

This work was supported by National Natural Science Foundation of China (61925201,62132001), Zhejiang Lab (NO.2022NB0AB05) and CCF-Tencent Open Research Fund.

References

[1] Hangbo Bao, Li Dong, Songhao Piao, and Furu Wei. Beit: Bert pre-training of image transformers. *arXiv preprint arXiv:2106.08254*, 2021. 3

[2] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020. 3

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

[4] Chaoqi Chen, Zebiao Zheng, Yue Huang, Xinghao Ding, and Yizhou Yu. I3net: Implicit instance-invariant network for adapting one-stage object detectors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12576–12585, 2021. 2

[5] Meilin Chen, Weijie Chen, Shicai Yang, Jie Song, Xinchao Wang, Lei Zhang, Yunfeng Yan, Donglian Qi, Yueting Zhuang, Di Xie, et al. Learning domain adaptive object detection with probabilistic teacher. In *International Conference on Machine Learning*, pages 3040–3055. PMLR, 2022. 1, 2, 6, 7

[6] Mark Chen, Alec Radford, Rewon Child, Jeffrey Wu, Heewoo Jun, David Luan, and Ilya Sutskever. Generative pretraining from pixels. In *International conference on machine learning*, pages 1691–1703. PMLR, 2020. 3

[7] Qingchao Chen and Yang Liu. Structure-aware feature fusion for unsupervised domain adaptation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 10567–10574, 2020. 2

[8] Qingchao Chen, Yang Liu, Zhaowen Wang, Ian Wassell, and Kevin Chetty. Re-weighted adversarial adaptation network for unsupervised domain adaptation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7976–7985, 2018. 2

[9] 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. 1, 2, 6, 7

[10] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3213–3223, 2016. 6

[11] Zhigang Dai, Bolun Cai, Yugeng Lin, and Junying Chen. Up-detr: Unsupervised pre-training for object detection with transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1601–1610, 2021. 3

[12] Jinhong Deng, Wen Li, Yuhua Chen, and Lixin Duan. Unbiased mean teacher for cross-domain object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4091–4101, 2021. 2, 6

[13] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, 2019. 3

[14] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2020. 3, 4

[15] Yaroslav Ganin and Victor Lempitsky. Unsupervised domain adaptation by backpropagation. In *International conference on machine learning*, pages 1180–1189. PMLR, 2015. 3

[16] Golnaz Ghiasi, Tsung-Yi Lin, and Quoc V Le. Dropblock: A regularization method for convolutional networks. *Advances in neural information processing systems*, 31, 2018. 3

[17] Kaixiong Gong, Shuang Li, Shugang Li, Rui Zhang, Chi Harold Liu, and Qiang Chen. Improving transferability for domain adaptive detection transformers. *arXiv preprint arXiv:2204.14195*, 2022. 2, 6, 7

[18] Song Han, Jeff Pool, Sharan Narang, Huizi Mao, Enhao Gong, Shijian Tang, Erich Elsen, Peter Vajda, Manohar Paluri, John Tran, et al. Dsd: Dense-sparse-dense training for deep neural networks. *arXiv preprint arXiv:1607.04381*, 2016. 2, 3

[19] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16000–16009, 2022. 2, 3, 5

[20] Mengzhe He, Yali Wang, Jiaxi Wu, Yiru Wang, Hanqing Li, Bo Li, Weihao Gan, Wei Wu, and Yu Qiao. Cross domain object detection by target-perceived dual branch distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9570–9580, 2022. 1, 2, 6

[21] Zhenwei He and Lei Zhang. Multi-adversarial faster-rcnn for unrestricted object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6668–6677, 2019. 2

[22] Cheng-Chun Hsu, Yi-Hsuan Tsai, Yen-Yu Lin, and Ming-Hsuan Yang. Every pixel matters: Center-aware feature alignment for domain adaptive object detector. In *European Conference on Computer Vision*, pages 733–748. Springer, 2020. 6, 7

[23] Gao Huang, Yixuan Li, Geoff Pleiss, Zhuang Liu, John E Hopcroft, and Kilian Q Weinberger. Snapshot ensembles: Train 1, get m for free. *arXiv preprint arXiv:1704.00109*, 2017. 3

[24] Wei-Jie Huang, Yu-Lin Lu, Shih-Yao Lin, Yusheng Xie, and Yen-Yu Lin. Aqt: Adversarial query transformers for domain

adaptive object detection. In *31st International Joint Conference on Artificial Intelligence, IJCAI 2022*, pages 972–979. International Joint Conferences on Artificial Intelligence, 2022. 2, 6, 7

[25] Junguang Jiang, Baixu Chen, Jianmin Wang, and Mingsheng Long. Decoupled adaptation for cross-domain object detection. In *International Conference on Learning Representations*, 2021. 6

[26] Matthew Johnson-Roberson, Charles Barto, Rounak Mehta, Sharath Nittur Sridhar, Karl Rosaen, and Ram Vasudevan. Driving in the matrix: Can virtual worlds replace human-generated annotations for real world tasks? *arXiv preprint arXiv:1610.01983*, 2016. 6

[27] 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/CVF Conference on Computer Vision and Pattern Recognition*, pages 12456–12465, 2019. 2

[28] Wuyang Li, Xinyu Liu, and Yixuan Yuan. Sigma: Semantic-complete graph matching for domain adaptive object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5291–5300, 2022. 2, 6

[29] Yu-Jhe Li, Xiaoliang Dai, Chih-Yao Ma, Yen-Cheng Liu, Kan Chen, Bichen Wu, Zijian He, Kris Kitani, and Peter Vajda. Cross-domain adaptive teacher for object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7581–7590, 2022. 1, 2, 6, 7

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

[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. 2

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

[33] Muhammad Akhtar Munir, Muhammad Haris Khan, M Sarfraz, and Mohsen Ali. Ssal: Synergizing between self-training and adversarial learning for domain adaptive object detection. *Advances in Neural Information Processing Systems*, 34:22770–22782, 2021. 6, 7

[34] Mengxue Qu, Yu Wu, Wu Liu, Qiqi Gong, Xiaodan Liang, Olga Russakovsky, Yao Zhao, and Yunchao Wei. Siri: A simple selective retraining mechanism for transformer-based visual grounding. In *European Conference on Computer Vision*, pages 546–562. Springer, 2022. 2, 3

[35] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28, 2015. 1, 2, 6, 7

[36] Farzaneh Rezaeianaran, Rakshith Shetty, Rahaf Aljundi, Daniel Olmeda Reino, Shanshan Zhang, and Bernt Schiele. Seeking similarities over differences: Similarity-based domain alignment for adaptive object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9204–9213, 2021. 2, 7

[37] Christos Sakaridis, Dengxin Dai, and Luc Van Gool. Semantic foggy scene understanding with synthetic data. *International Journal of Computer Vision*, 126(9):973–992, 2018. 6

[38] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 15(1):1929–1958, 2014. 3

[39] Kun Tian, Chenghao Zhang, Ying Wang, Shiming Xiang, and Chunhong Pan. Knowledge mining and transferring for domain adaptive object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9133–9142, 2021. 7

[40] Zhi Tian, Chunhua Shen, Hao Chen, and Tong He. Fcos: Fully convolutional one-stage object detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9627–9636, 2019. 1, 2, 6, 7

[41] Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training. *arXiv preprint arXiv:2203.12602*, 2022. 2, 3

[42] Pascal Vincent, Hugo Larochelle, Yoshua Bengio, and Pierre-Antoine Manzagol. Extracting and composing robust features with denoising autoencoders. In *Proceedings of the 25th international conference on Machine learning*, pages 1096–1103, 2008. 3

[43] Vibashan Vs, Vikram Gupta, Poojan Oza, Vishwanath A Sindagi, and Vishal M Patel. Mega-cda: Memory guided attention for category-aware unsupervised domain adaptive object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4516–4526, 2021. 7

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

[45] Chang-Dong Xu, Xing-Ran Zhao, Xin Jin, and Xiu-Shen Wei. Exploring categorical regularization for domain adaptive object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11724–11733, 2020. 7

[46] 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. 2, 3

[47] Fisher Yu, Haofeng Chen, Xin Wang, Wenqi Xian, Yingying Chen, Fangchen Liu, Vashisht Madhavan, and Trevor Darrell. Bdd100k: A diverse driving dataset for heterogeneous multitask learning. In *Proceedings of the IEEE/CVF con-*

ference on computer vision and pattern recognition, pages 2636–2645, 2020. 6

[48] Jinze Yu, Jiaming Liu, Xiaobao Wei, Haoyi Zhou, Yohei Nakata, Denis Gudovskiy, Tomoyuki Okuno, Jianxin Li, Kurt Keutzer, and Shanghang Zhang. Cross-domain object detection with mean-teacher transformer. *arXiv preprint arXiv:2205.01643*, 2022. 1, 2, 3, 6, 7

[49] Dan Zhang, Jingjing Li, Lin Xiong, Lan Lin, Mao Ye, and Shangming Yang. Cycle-consistent domain adaptive faster rcnn. *IEEE Access*, 7:123903–123911, 2019. 1, 2

[50] Liang Zhao and Limin Wang. Task-specific inconsistency alignment for domain adaptive object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14217–14226, 2022. 6

[51] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable detr: Deformable transformers for end-to-end object detection. In *International Conference on Learning Representations*, 2020. 1, 2, 3, 6, 7