

DA-F2F: Domain-Adaptive Object Detection with Feature-to-Feature Modulation and Alignment

HoTaek Oh, Hee-Jun Kim, and Hyo-Jun Lee *

Kangwon National University
{peterqwer123, bnkhj2507, gywns6298}@kangwon.ac.kr

Abstract. Domain adaptive object detection (DAOD) is essential for deploying reliable detectors in real-world environments with inevitable domain shifts. While recent approaches rely on image-to-image translation, they often suffer from instability inherent in pixel-level domain transformation. To address these limitations, we propose DA-F2F, a novel framework that directly aligns domains within the feature space. DA-F2F introduces a style-aware feature modulation (SFM) module that extracts style statistics from the target domain to dynamically modulate source representations. Furthermore, to improve adversarial learning, we propose a soft-weighted proposal alignment (SPA) module that stably separates foreground and background regions in the target domain. Extensive experiments demonstrate that DA-F2F significantly outperforms existing state-of-the-art methods across diverse adaptation scenarios, achieving 60.5% mAP on clear-to-adverse weather (Foggy Cityscapes), 46.9% mAP on small-to-large-scale datasets (BDD100K), and 69.9% mAP on synthetic-to-real adaptation (SIM10K). Code is available at <https://github.com/ohhotaek/DA-F2F>.

Keywords: Domain Adaptive Object Detection · Feature-to-Feature Modulation · Region-Separable Adversarial Learning

1 Introduction

Object detection is a fundamental task in computer vision, essential for developing diverse downstream applications such as autonomous driving, surveillance, and robotics [12, 29, 35, 36]. While modern supervised detectors achieve remarkable performance when trained on large-scale annotated datasets, these models often suffer from severe performance degradation when deployed in out-of-distribution environments. Such domain shifts, arising from variations in lighting, weather, or camera configurations, significantly hinder the real-world scalability of object detectors. Since obtaining high-quality labels for every new target domain is labor-intensive and often impractical, developing a robust framework that can effectively adapt to unseen domains remains a pivotal challenge.

To address these limitations, domain adaptive object detection (DAOD) has emerged as a promising solution, aiming to generalize models trained on a labeled source domain to an unlabeled target domain. Recently, the mean teacher

* Corresponding author.

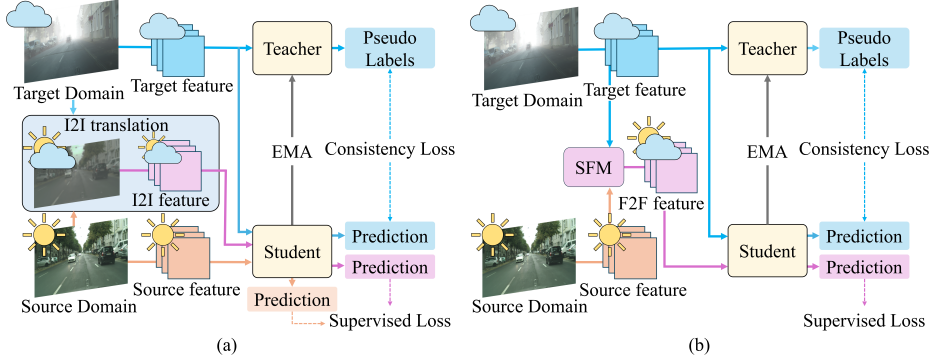


Fig. 1: Conceptual comparison between conventional image-to-image (I2I) translation and the proposed feature-to-feature (F2F) translation. (a) I2I translation methods synthesize target-like images from the source domain to utilize them as auxiliary supervised data for the student detector. (b) The proposed DA-F2F achieves feature-to-feature modulation, which directly transforms source domain representations into the target domain within the latent feature space.

paradigm [38] has achieved compelling success in DAOD tasks [1, 4, 10, 38, 40, 48]. This framework optimizes the student detector on the target domain by jointly minimizing a standard supervised loss on the labeled source domain and a consistency loss between the student’s predictions and the pseudo-labels generated by the teacher. Subsequently, the teacher is updated through an exponential moving average (EMA) [28, 31] of the student parameters to ensure high-quality and stable predictions. However, the teacher in this framework often generates unstable pseudo-labels, especially under severe visual discrepancies between the source and target domains. This occurs because the EMA-updated teacher inherently relies on the student’s parameters, which are inevitably biased towards the source domain due to the supervised loss. Consequently, these noisy pseudo-labels affect the training process, hindering steady performance optimization.

Alternatively, several straightforward methods [9, 32, 40, 50] employ image-to-image translation to synthesize a target-like domain from source images, as shown in Fig. 1(a). Since these target-like images originate from the source domain, they retain their corresponding ground-truth labels and can be directly utilized as auxiliary supervised training data for the student model. However, these approaches typically rely on off-the-shelf style transfer networks [33, 51], introducing significant computational overhead. Furthermore, if the quality of the synthesized image is unstable or exhibits pixel-level artifacts, the overall detection performance can be severely degraded [50]. Therefore, a more stable and efficient domain transformation approach is needed to avoid the inherent risks of direct pixel reconstruction while achieving computational efficiency.

In this study, we propose a novel domain adaptive object detection framework, DA-F2F, faithfully designed to address the instability of pixel-level domain transformation, as illustrated in Fig. 1(b). Rather than relying on explicit image translation, the proposed DA-F2F achieves domain alignment through a feature-

to-feature modulation approach using a style-aware feature modulation (SFM) module. The proposed SFM extracts style statistics from target features to dynamically generate style modulation parameters, which are subsequently applied to the source features to effectively align their representations with the target style distribution.

Another prominent DAOD approach is adversarial learning [5, 6, 16, 42, 47], which employs a domain discriminator on the image features and induces the network to deceive this discriminator, thereby alleviating feature discrepancy between domains. Some recent methods [6, 16, 42, 47] advance this paradigm by distinguishing the foreground and background regions within the image features to achieve more precise alignment. We also adopt this distinguishing strategy, while addressing the unstable separation of foreground and background in the target domain using a soft-weighted proposal alignment (SPA). Extensive experiments demonstrate that the proposed DA-F2F outperforms the other state-of-the-arts in multiple adaptation scenarios, including the weather adaptation (Cityscapes [7] $\rightarrow$ Foggy Cityscapes [37]), the small-to-large scale dataset adaptation (Cityscapes [7] $\rightarrow$ BDD100K [44]), and the synthetic-to-real adaptation (Sim10k [20] $\rightarrow$ Cityscapes [37]). To summarize, the main contributions of our work are threefold:

- We develop the style-aware feature modulation that effectively implements feature to feature modulation to transform source domain representations into the target domain within the feature space.
- We also introduce the soft-weighted proposal alignment that advances adversarial learning by stably separating foreground and background regions in the target domain.
- Extensive experiments across various domain shift scenarios demonstrate that the proposed framework achieves outstanding performance in DAOD.

2 Related Work

2.1 Mean Teacher Self-Training Methods

Domain adaptive object detection (DAOD) aims to mitigate the performance degradation of object detectors across domains with different data distributions. Recently, mean teacher self-training has achieved compelling success in DAOD tasks [1, 4, 10, 38, 40, 48]. Specifically, AT [28] uses a self-training paradigm with weak-strong data augmentation, yielding a noticeable increase in accuracy. HT [10] improves pseudo-label quality by regularizing the consistency between classification and localization, and PT [4] performs reliability-based selection by quantifying pseudo-label uncertainty. CMT [1] strengthens representation robustness through object-level contrastive learning, and MRT [48] refines feature representations by integrating masked reconstruction with selective retraining.

Mean Teacher Methods with Image-to-Image Translation. To address the considerable model bias inherent in mean teacher self-training methods, several approaches [9, 32, 40, 50] employ image-to-image translation to synthesize

a target-like domain from source images. UMT [9] utilizes image translation by cross-domain distillation for the teacher model and pixel-level adaptation for the student model. DDT [40] creates source-style and target-style images from the target and source samples using the CycleGAN [51], and initializes the teacher-student framework with labeled source and target-style images. However, these approaches typically rely on off-the-shelf style transfer networks [33, 51], introducing significant computational overhead and performance degradation caused by unstable pixel-level artifacts. In contrast, we utilize a more efficient and stable domain transformation approach based on feature-to-feature modulation.

2.2 DAOD with Adversarial Learning

Another representative domain alignment strategy in DAOD is adversarial learning. As a pioneering work, DA-Faster [5] learns domain-invariant representations by inverting the gradient direction during backpropagation to confuse the domain discriminator at both image and instance levels. Subsequently, OADA [43] mitigates domain discrepancies at the localization level by leveraging bounding box regression offsets. Furthermore, methods such as SIGMA [26], CIGAR [30], and CAT [21] introduce relation-based approaches with adversarial learning to alleviate class-conditional distribution shifts.

Adversarial Learning with Transformer. Several methods bring adversarial learning into the transformer-based detector. O2Net [13], AQT [18], and SFA [41] adopt adversarial learning at the object, query, and sequence levels, respectively. BiADT [15] explicitly disentangles domain-invariant and domain-specific tokens using adversarial learning. DA-DETR [46] executes alignment by integrating CNN and Transformer representations, while DATR [3] further introduces class-wise prototype alignment and dataset-level adaptation within a DETR-based framework to enhance instance-level cross-domain feature consistency. Recently, the research paradigm has expanded to leverage vision-language models (VLM). For instance, DA-Pro [22] applies domain-aware prompt tuning based on a frozen VLM to learn a domain-adaptive detection head. SEEN-DA [25] isolates domain-specific semantic representations through a semantic entropy-based attention mechanism. Furthermore, DA-Ada [23] inserts domain-aware adapters into a frozen VLM backbone to jointly learn domain-invariant and domain-specific knowledge.

Region Separable Adversarial Learning. To overcome the limitations of global alignment, recent methods [6, 16, 42, 47] advance the adversarial learning paradigm by explicitly distinguishing the foreground and background regions. HMT [42] extends this separable adversarial learning within the Transformer-based DETR architecture by complementing the pseudo-labeling strategy and matching mechanisms. Similarly, DADAOD [16] and REACT [24] propose selective alignment strategies to refine the adaptation process. We also adopt a region-separated distinguishing strategy, while addressing the unstable separation of foreground and background in the target domain using a soft-weighted proposal alignment.

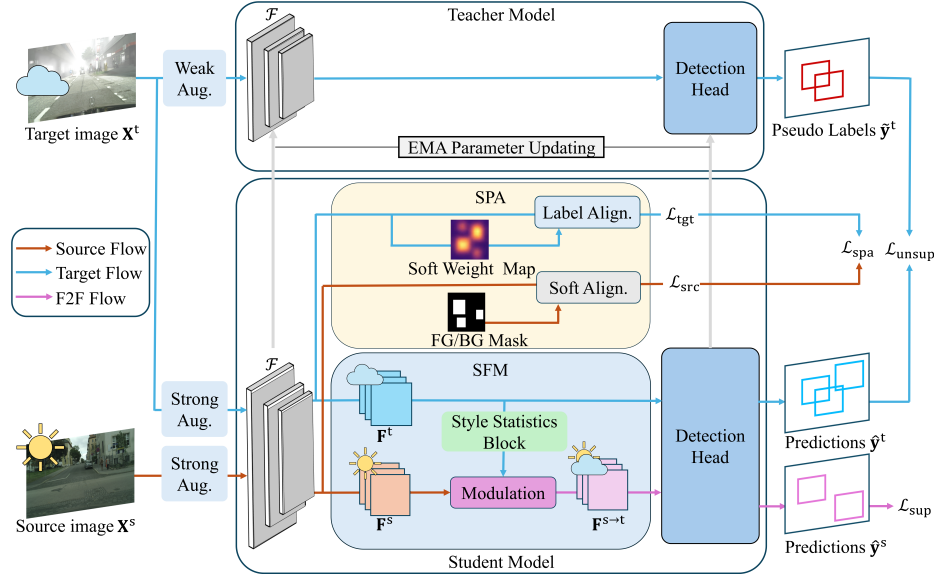


Fig. 2: Overview of proposed DA-F2F framework. DA-F2F adopts a teacher-student architecture with an EMA-updated teacher. The style-aware feature modulation (SFM) module transforms source features into target-aligned representations by utilizing the style statistics extracted from target features. The soft-weighted proposal alignment (SPA) module facilitates region-separated adversarial learning for both source and target domains, specifically employing a soft-weighted region separation strategy to robustly distinguish foreground and background regions within the target features.

3 Method

3.1 Preliminaries

In the DAOD framework, we are given N_s labeled source domain samples $\mathcal{S}^{\text{src}} = \{(\mathbf{X}_i^s, \mathbf{y}_i^s)\}_{i=1}^{N_s}$ and N_t unlabeled target domain samples $\mathcal{S}^{\text{tar}} = \{\mathbf{X}_i^t\}_{i=1}^{N_t}$, where $\mathbf{X}$ denotes an image and $\mathbf{y} = (\mathbf{b}, \mathbf{o})$ represents the annotation labels including the bounding box coordinates $\mathbf{b}$ and the object class $\mathbf{o}$. The primary goal of DAOD is to train a domain invariant detector by leveraging both $\mathcal{S}^{\text{src}}$ and $\mathcal{S}^{\text{tar}}$.

As shown in Fig. 2, the proposed DA-F2F adopts the teacher-student self training paradigm. In this paradigm, the teacher and student share the backbone and detector network structure. The teacher processes each weakly augmented target image $\mathbf{X}^t$ to generate pseudo-labels $\hat{\mathbf{y}}^t$, while the student is optimized based on the supervision provided by the ground-truth labels of the source domain $\mathcal{S}^{\text{src}}$ and the pseudo-labels of the target domain samples. Then, the teacher network is updated via an exponential moving average (EMA) [28, 31] of the student parameters:

$$\theta_{\text{tea}} \leftarrow \alpha \theta_{\text{tea}} + (1 - \alpha) \theta_{\text{stu}}, \quad (1)$$

where θ_{tea} and θ_{stu} denote the parameters of the teacher and student, respectively. $\alpha \in [0, 1)$ is the momentum hyperparameter of the EMA, which is set to 0.9996. Through this iterative process, the student progressively distills domain-invariant features, eventually emerging as a robust detector capable of generalizing effectively across both source and target domains.

3.2 Framework Overview

Fig. 2 illustrates the overall framework of DA-F2F. During the training process, the student model in DA-F2F processes pairs of a source image $\mathbf{X}^s$ and a target image $\mathbf{X}^t$. DA-F2F then employs backbone networks $\mathcal{F}$ to extract source features $\mathbf{F}^s \in \mathbb{R}^{H \times W \times C}$ and target features $\mathbf{F}^t \in \mathbb{R}^{H \times W \times C}$, where (H, W) denotes the spatial size of the feature map and C represents the channel dimension. Note that we describe the process using a single scale of backbone features for clarity, although the framework operates across multiple scales in practice.

To integrate the target representation into the source features, the style-aware feature modulation (Section 3.3) transforms the source features $\mathbf{F}^s$ into target-aligned features $\mathbf{F}^{s \rightarrow t} \in \mathbb{R}^{H \times W \times C}$ using the style statistics of the target features $\mathbf{F}^t$. This feature-to-feature modulation transfers domain-specific attributes directly within the feature space, which alleviates the mode collapse and computational overhead typically associated with high-dimensional image spaces. The soft-weighted proposal alignment (Section 3.4) employs region-separated adversarial learning for both source and target features by distinguishing between foreground and background regions within each feature map. While partitioning these regions in source features is straightforward due to the available ground truth annotations, isolating each region within the target features introduces instability due to its dependency on noisy pseudo-labels generated by a teacher model. Therefore, the proposed soft-weighted proposal alignment adopts a soft weighted region separation strategy for the target features.

Finally, the student estimates predictions $\hat{\mathbf{y}}^s$ for the source domain sample and $\hat{\mathbf{y}}^t$ for the target domain sample from the target-aligned source features $\mathbf{F}^{s \rightarrow t}$ and the target features $\mathbf{F}^t$, respectively. Concurrently, the teacher estimates the pseudo-labels $\tilde{\mathbf{y}}^t$ from the target image $\mathbf{X}^t$. Given the final predictions $\hat{\mathbf{y}}^s, \hat{\mathbf{y}}^t$ and labels $\mathbf{y}^s, \tilde{\mathbf{y}}^t$, the student and teacher are iteratively optimized based on the teacher-student training paradigm. Through these proposed feature alignment strategies, DA-F2F achieves highly aligned representations for robust DAOD.

3.3 Style-Aware Feature Modulation

Given pairs of source features $\mathbf{F}^s$ and target features $\mathbf{F}^t$, the style aware feature modulation (SFM) directly transfers domain specific attributes from the target features $\mathbf{F}^t$ to the source features $\mathbf{F}^s$, consequently generating the target aligned source features $\mathbf{F}^{s \rightarrow t}$. Fig. 3(a) illustrates the detailed process of the SFM, which consists of a style statistic block and a modulation block.

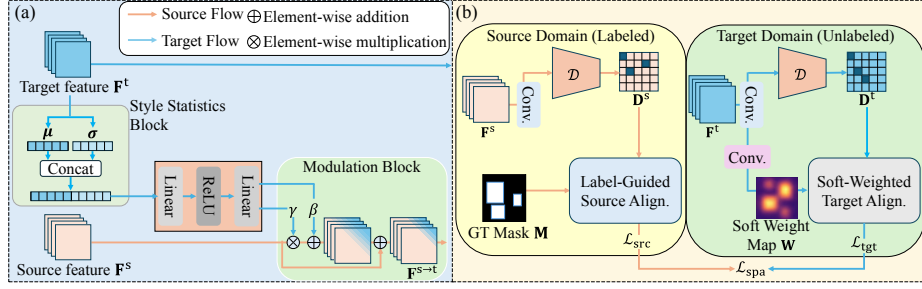


Fig. 3: (a) SFM extracts channel-wise style statistics from target features and maps them into affine parameters γ and β via an MLP. Given γ and β , SFM modulates source features into target-aligned source features $F^{s \rightarrow t}$. (b) SPA performs region-aware adversarial learning on source and target features, employing a label-guided alignment for the source features and a soft-weighted alignment for the unlabeled target features.

Style Statistic Block. In deep feature representations, it is well-established that channel-wise statistics effectively encapsulate domain-specific style characteristics, such as illumination, weather conditions, and textures [19, 27, 34]. Motivated by this, the style statistic block derives the affine transformation parameters required to modulate the source features based on the style statistics of the target domain. Specifically, the block first computes the style statistics, specifically the channel wise mean $\mu \in \mathbb{R}^C$ and standard deviation $\sigma \in \mathbb{R}^C$ from the target features F^t , followed by the calculation of their c -th elements μ_c and σ_c :

$$\mu_c = \frac{1}{HW} \sum_{h=1}^H \sum_{w=1}^W F_{h,w,c}^t, \quad \sigma_c = \sqrt{\frac{1}{HW} \sum_{h=1}^H \sum_{w=1}^W (F_{h,w,c}^t - \mu_c)^2 + \epsilon}, \quad (2)$$

where $F_{h,w,c}^t$ denotes the feature value at spatial location (h, w) for the c -th channel, and ϵ is a small constant. Subsequently, the style statistic block concatenates the two statistics $[\mu, \sigma] \in \mathbb{R}^{2C}$ and transforms this concatenated vector into an affine translation parameter $\beta \in \mathbb{R}^C$ and a scaling offset $\Delta_\gamma \in \mathbb{R}^C$ through a two layer MLP with ReLU activation. The scaling offset is transformed into the final affine scaling parameter via $\gamma = 1 + \Delta_\gamma$. These affine transformation parameters, β and γ , subsequently modulate the source features in the subsequent modulation block.

Modulation Block. The modulation block aims to align the source features F^s with the target domain using the derived affine transformation parameters, β and γ . We specifically adopt this affine transformation because it effectively adjusts the style distributions while strictly preserving the underlying contextual information of the source domain, which is essential for accurate object detection. Therefore, the modulation block yields the c -th channel of the target aligned

features $\mathbf{F}_c^{s \rightarrow t}$ as follows:

$$\tilde{\mathbf{F}}_c^{s \rightarrow t} = \gamma_c \mathbf{F}_c^s + \beta_c \quad (3)$$

$$\mathbf{F}_c^{s \rightarrow t} = (1 - w) \mathbf{F}_c^s + w \tilde{\mathbf{F}}_c^{s \rightarrow t}, \quad (4)$$

where w is a weighting parameter to balance the modulation. These target aligned source representations encourage the network to extract domain invariant features, thereby boosting the overall robustness of the DAOD framework.

3.4 Soft-Weighted Proposal Alignment

Adversarial learning [5, 6, 16, 42, 47] is a common DAOD paradigm to align features across the source and target domains. Several studies [6, 16, 42, 47] have revealed that separating foreground and background regions during adversarial learning achieves more robust feature alignments. Even though the separation of these regions for the source features is guaranteed by ground-truth labels, dividing the target features based on imperfect pseudo-labels inevitably leads to inaccurate region assignments. To address this, we introduce the soft-weighted proposal alignment (SPA), which replaces discrete region assignments with a continuous soft weighting strategy as illustrated in Fig. 3(b). Specifically, it utilizes two different region separation strategies for the source and target features.

Label-Guided Source Alignment. To align the source domain representations, we first forward the source features $\mathbf{F}^s$ into a spatial domain discriminator $\mathcal{D}$ consisting of sequential convolutional layers. The discriminator processes the source features to predict a domain prediction map $\mathbf{D}^s = \mathcal{D}(\mathbf{F}^s)$, where each spatial location of $\mathbf{D}^s \in [0, 1]^{H \times W}$ indicates the probability of the feature belonging to the source domain. Instead of applying a uniform alignment objective across the entire spatial resolution of $\mathbf{D}^s$, we explicitly separate the foreground and background regions using ground-truth labels. Specifically, we define a binary spatial mask $\mathbf{M} \in \{0, 1\}^{H \times W}$, where values are assigned to 1 for locations situated within the ground-truth bounding boxes and 0 otherwise. Finally, the separable alignment objective can be formulated as follows:

$$\mathcal{L}_{\text{fg}} = -\frac{1}{\|\mathbf{M}\|_1} \sum_{h=1}^H \sum_{w=1}^W \mathbf{M}_{h,w} \log \mathbf{D}_{h,w}^s \quad (5)$$

$$\mathcal{L}_{\text{bg}} = -\frac{1}{\|\mathbf{1}_{HW} - \mathbf{M}\|_1} \sum_{h=1}^H \sum_{w=1}^W (1 - \mathbf{M}_{h,w}) \log \mathbf{D}_{h,w}^s \quad (6)$$

$$\mathcal{L}_{\text{src}} = \lambda_{\text{fg}} \mathcal{L}_{\text{fg}} + \lambda_{\text{bg}} \mathcal{L}_{\text{bg}} \quad (7)$$

where $\mathbf{1}_{HW} \in \{1\}^{H \times W}$ is a matrix of ones, and λ_{fg} and λ_{bg} are hyper parameters balancing the relative importance of the foreground and background alignments. We set λ_{fg} to 0.8 and λ_{bg} to 0.05 to enable the adversarial learning to focus more on foreground regions, following [6, 16, 42, 47].

Soft-Weighted Target Alignment. The lack of ground-truth annotations in the target domain makes it challenging to perform discrete region-based alignment. To address this, we introduce a continuous soft-weighting strategy as illustrated in Fig. 3(b). First, the target features $\mathbf{F}^t$ are processed through sequential convolutional layers to generate a spatial proposal map $\mathbf{P} \in [0, 1]^{H \times W}$, optimized via the standard region proposal loss [36] using pseudo-labels. To mitigate the extreme weight disparity between foreground and background regions, we rescale the proposal map $\mathbf{P}$ into a bounded weight map $\mathbf{W} \in [\tau_{\min}, \tau_{\max}]^{H \times W}$ by

$$\mathbf{W} = \tau_{\min} \mathbf{1}_{HW} + (\tau_{\max} - \tau_{\min}) \mathbf{P}. \quad (8)$$

We set the hyper-parameters $\tau_{\min}$ and $\tau_{\max}$ to 0.05 and 0.8, respectively. Finally, the target features $\mathbf{F}^t$ are fed into the shared spatial domain discriminator $\mathcal{D}$ to obtain the domain probability map $\mathbf{D}^t = \mathcal{D}(\mathbf{F}^t)$. Then, the soft-weighted alignment objective can be formulated by:

$$\mathcal{L}_{\text{tgt}} = -\frac{1}{HW} \sum_{h=1}^H \sum_{w=1}^W \mathbf{W}_{h,w} \log(1 - \mathbf{D}_{h,w}^t), \quad (9)$$

where the integration of $\mathbf{W}$ forces the adversarial learning to focus more on proposal regions while elegantly suppressing the negative impact of the background. The final adversarial alignment objective for the proposed SPA module is then defined as:

$$\mathcal{L}_{\text{spa}} = \max_{\mathcal{F}} \min_{\mathcal{D}} (\mathcal{L}_{\text{src}} + \mathcal{L}_{\text{tgt}}), \quad (10)$$

where $\mathcal{F}$ denotes the backbone networks.

3.5 Overall Training Objective

The overall training objective of our proposed DA-F2F integrates the standard DAOD losses with our adversarial alignment loss. The total loss $\mathcal{L}_{\text{total}}$ to optimize the entire network is formulated as follows:

$$\mathcal{L}_{\text{total}} = \lambda_{\text{sup}} \mathcal{L}_{\text{sup}} + \lambda_{\text{unsup}} \mathcal{L}_{\text{unsup}} + \lambda_{\text{rpn}} \mathcal{L}_{\text{rpn}} + \lambda_{\text{spa}} \mathcal{L}_{\text{spa}}, \quad (11)$$

where $\mathcal{L}_{\text{sup}}$ represents the supervised detection loss between the ground-truth labels $\mathbf{y}^s$ and the student predictions $\hat{\mathbf{y}}^s$ for the source domain samples, while $\mathcal{L}_{\text{unsup}}$ represents the unsupervised detection loss between the pseudo-labels $\tilde{\mathbf{y}}^t$ from the teacher and the student predictions $\hat{\mathbf{y}}^t$ for the target domain samples. $\mathcal{L}_{\text{rpn}}$ is the region proposal loss [36] used to optimize the soft target proposal map $\mathbf{P}$ of SPA, and $\mathcal{L}_{\text{spa}}$ denotes our proposed adversarial alignment loss. We empirically set the scaling weights λ_{sup} , λ_{unsup} , λ_{rpn} , and λ_{spa} to 1, 1, 1, and 0.01, respectively. Through this joint optimization, the feature extractor learns robust representations that are both highly discriminative for the detection task and invariant across the source and target domains.

4 Experiments

4.1 Datasets

Following state-of-the-art DAOD methods [3, 16, 24, 42], we evaluate the proposed DA-F2F on three widely-used benchmarks. Specifically, we perform adaptation experiments across three common scenarios: (1) weather adaptation (Cityscapes $\rightarrow$ Foggy Cityscapes), (2) small-to-large scale dataset adaptation (Cityscapes $\rightarrow$ BDD100K-daytime), and (3) synthetic-to-real adaptation (Sim10k $\rightarrow$ Cityscapes). For all experiments, we employ the mean Average Precision (mAP) with an intersection-over-union (IoU) threshold of 0.5 as the primary evaluation metric. **Cityscapes** [7] is a real-world dataset comprising urban street scenes captured across 50 different cities under clear weather conditions. It contains 2,975 training and 500 validation images. Consistent with existing DAOD methods, we convert the officially provided semantic segmentation labels into bounding box annotations for our DAOD task.

Foggy Cityscapes [37] is a synthetic dataset rendered from Cityscapes images and depth maps, simulating three fog densities (0.005, 0.01, 0.02). It inherits the identical data split and bounding box annotations as the original Cityscapes. For a fair comparison with previous works, we utilize the highest fog density (0.02) as our target domain for evaluation.

BDD100K-daytime [44] is a subset of the large-scale BDD100K autonomous driving dataset, specifically designed for daytime scenarios. It contains 36,728 training and 5,258 validation images, providing a significantly larger and more diverse environment compared to Cityscapes.

Sim10k [20] is a synthetic dataset containing 10,000 training images rendered from the Grand Theft Auto V (GTA-V) game engine. It provides 58,701 car bounding box annotations, which are commonly used to evaluate the synthetic-to-real adaptation capability of DAOD models.

4.2 Implementation Details

Network Structure. We adopt Faster R-CNN [36] as the base detector, employing ResNet-50 [14] and a Feature Pyramid Network (FPN) [29] as the backbone. The backbone is initialized with ImageNet [8] pre-trained weights, and its first two stages are frozen to ensure training stability. The channel dimension of the FPN features is set to $C = 256$. The input images are resized to a resolution of 1024×2048 . Additionally, the modulation weight w in Eq. (4) is set to 0.3.

Training. DA-F2F is optimized for a total of 40k iterations, comprising 10k iterations for the burn-in stage and 30k iterations for teacher-student mutual learning. During the burn-in stage, the network is trained solely on the labeled source domain dataset, while the mutual learning stage utilizes both labeled source and unlabeled target images. We employ the stochastic gradient descent (SGD) optimizer with a momentum of 0.9. The initial learning rate is set to 0.01 and gradually decays towards the final iteration. The batch size is set to 32, consisting of 16 labeled source images and 16 unlabeled target images. During self-training, we apply common weak augmentation techniques in DAOD,

Table 1: Experimental results (%) on the weather adaptation scenario: Cityscapes $\rightarrow$ Foggy Cityscapes. We highlight the best results as a **bold**.

Cityscapes $\rightarrow$ Foggy Cityscapes										
Method	Detector	person	rider	car	truck	bus	train	mcycle	bicycle	mAP
SIGMA [26]	FCOS	44.0	43.9	60.3	31.6	50.4	51.5	31.7	40.6	44.2
OADA [43]	FCOS	47.8	46.5	62.9	32.1	48.5	50.9	34.3	39.8	45.4
HT [10]	FCOS	52.1	55.8	67.5	32.7	55.9	49.1	40.1	50.3	50.4
DA-DETR [46]	Transformer	49.9	50.0	63.1	24.0	45.8	37.5	31.6	46.3	43.5
O2Net [13]	Transformer	48.7	51.5	63.6	31.1	47.6	47.8	38.0	45.9	46.8
AQT [18]	Transformer	49.3	52.3	64.4	27.7	53.7	46.5	36.0	46.4	47.1
BiADT [15]	Transformer	52.2	58.9	69.2	31.7	55.0	45.1	42.6	51.3	50.8
MRT [48]	Transformer	52.8	51.7	68.7	35.9	58.1	54.5	41.0	47.1	51.2
DATR [3]	Transformer	61.6	60.4	74.3	35.7	60.3	35.4	43.6	55.9	53.4
HMT [42]	Transformer	57.0	62.8	74.1	37.3	61.0	57.7	44.4	57.5	56.5
DA-Pro [22]	RegionCLIP	55.4	62.9	70.9	40.3	63.4	54.0	42.3	58.0	55.9
SEEN-DA [25]	RegionCLIP	58.5	64.5	71.7	42.0	61.2	54.8	47.1	59.9	57.5
DA-Ada [23]	RegionCLIP	57.8	65.1	71.3	43.1	64.0	58.6	48.8	58.7	58.5
DA-Faster [5]	FRCNN	29.2	40.4	43.4	19.7	38.3	28.5	23.7	32.7	32.0
UMT [9]	FRCNN	33.0	46.7	48.6	34.1	56.5	46.8	30.4	37.3	41.7
PT [4]	FRCNN	43.2	52.4	63.4	33.4	56.6	37.8	41.3	48.7	47.1
CMT [1]	FRCNN	47.0	55.7	64.5	39.4	63.2	51.9	40.3	53.1	51.9
CAT [21]	FRCNN	44.6	57.1	63.7	40.8	66.0	49.7	44.9	53.0	52.5
REACT [24]	FRCNN	51.4	57.9	67.4	37.7	58.4	52.8	44.6	54.6	53.1
DDT [40]	FRCNN	60.9	64.9	75.6	36.3	60.6	53.7	42.3	59.3	56.7
DADAOD [16]	FRCNN	59.8	62.8	73.7	40.3	59.4	56.1	47.8	58.3	57.3
DA-F2F	FRCNN	60.9	63.0	76.3	46.2	62.8	64.3	52.8	58.2	60.5

including random contrast, brightness, saturation, grayscale, and blur. For the source images, strong augmentation consists of weak augmentations combined with random erasing [11]. For the target domain images, strong augmentation includes weak augmentations along with the Masking Image Consistency (MIC) [17] method. Pseudo-labels generated by the teacher model are filtered using a confidence threshold of 0.8. All processes in this study are implemented using four NVIDIA RTX A6000 GPUs.

4.3 Main Results

Tables 1, 2, and 3 compare the proposed DA-F2F with existing state-of-the-art DAOD methods across three different adaptation scenarios. The notations, FCOS, Transformer, RegionCLIP, and FRCNN indicate that the corresponding methods utilize FCOS [39], Transformer-based detectors [2, 45], RegionCLIP [49], and Faster R-CNN [36] as their teacher-student detector. For a fair comparison, all baseline scores are reported directly from their respective original papers.

Adaptation for Clear-to-Adverse Weather. Table 1 presents the quantitative comparisons of DA-F2F with other state-of-the-art methods on weather adaptation. Following the standard protocol [3, 16, 24, 42], we utilize the images and labels from the Cityscapes training set as the source domain, while the images from the Foggy Cityscapes training set serve as the target domain. The final evaluation is performed on the Foggy Cityscapes validation set. As shown

Table 2: Experimental results (%) on the small-to-large scale adaptation scenario: Cityscapes $\rightarrow$ BDD100k-daytime. We highlight the best results as a **bold**.

Cityscapes $\rightarrow$ BDD100k-daytime									
Method	Detector	person	rider	car	truck	bus	mcycle	bicycle	mAP
SIGMA [26]	FCOS	46.9	29.6	64.1	20.2	23.6	17.9	26.3	32.7
HT [10]	FCOS	53.4	40.4	63.5	27.4	30.6	28.2	38.0	40.2
AQT [18]	Transformer	38.2	33.0	58.4	17.3	18.4	16.9	23.5	29.4
O2Net [13]	Transformer	40.4	31.2	58.6	20.4	25.0	14.9	22.7	30.5
BiADT [15]	Transformer	42.0	34.5	59.9	17.2	19.2	17.8	24.4	32.7
MRT [48]	Transformer	48.4	30.9	63.7	24.7	25.5	20.2	22.6	33.7
HMT [42]	Transformer	52.6	42.5	67.8	33.7	36.3	28.6	33.7	42.2
DATR [3]	Transformer	58.5	42.8	73.4	26.9	39.9	24.2	37.3	43.3
DA-Faster [5]	FRCNN	28.9	27.4	44.2	19.1	18.0	14.2	22.4	24.9
PT [4]	FRCNN	40.5	39.9	52.7	25.8	33.8	23.0	28.8	34.9
REACT [24]	FRCNN	-	-	-	-	-	-	-	35.8
CAT [21]	FRCNN	44.6	41.5	61.2	31.4	34.6	24.4	31.7	38.5
DADAOD [16]	FRCNN	61.4	45.4	75.4	33.0	36.2	29.5	36.7	45.8
DA-F2F	FRCNN	64.9	50.0	78.6	33.9	38.4	35.2	39.2	48.6

in Table 1, the proposed DA-F2F significantly outperforms existing methods on weather adaptation with 60.5% mAP. Compared to the previous model with the best performance, DA-Ada [23], DA-F2F achieves a notable gain of +2.0% mAP. Furthermore, when compared to methods utilizing the same base detector, DA-F2F significantly surpasses DADAOD [16] by a large margin of 3.2% mAP.

Adaptation for Small-to-Large-Scale Dataset. Following the established settings, we adopt the Cityscapes training set as the source domain and the BDD100K-daytime training set as the target domain, with evaluation performed on its validation set. As presented in Table 2, DA-F2F consistently achieves the highest DAOD performance across various semantic classes. Consequently, DA-F2F attains the highest overall mAP of 48.6%, outperforming DADAOD [16] by a margin of +2.8% mAP.

Adaptation for Synthetic-to-Real. Consistent with previous studies, we employ the synthetic Sim10k dataset as the source domain and the real-world Cityscapes training set as the target domain. The performance is evaluated exclusively on the car category using the Cityscapes validation set. Table 3 demonstrates that the proposed DA-F2F still achieves the best results in the synthetic-to-real adaptation scenario, yielding a state-of-the-art mAP of 69.9%.

Qualitative Results. Fig. 4 presents qualitative comparisons between DA-F2F and DADAOD [16] across three adaptation scenarios. We generated the detection results of DADAOD following its official code and instructions. We observe that DA-F2F is more robust, consistently achieving high-quality detection results across all scenarios.

4.4 Ablation Study

We analyze the contributions of the individual components within the proposed DA-F2F framework through ablation experiments on the Cityscapes $\rightarrow$ Foggy

Table 3: Experimental results (%) on the synthetic-to-real adaptation scenario: Sim10K $\rightarrow$ Cityscapes solely on car category. We highlight the best results as a **bold**.

SIM10K $\rightarrow$ Cityscapes		
Method	Detector	mAP
SIGMA [26]	FCOS	53.7
OADA [43]	FCOS	59.2
HT [10]	FCOS	65.5
AQT [18]	Trans.	53.4
O2Net [13]	Trans.	54.1
DA-DETR [46]	Trans.	54.7
BiADT [15]	Trans.	55.8
MRT [48]	Trans.	62.0
DATR [3]	Trans.	64.3
HMT [42]	Trans.	66.4
DA-Pro [22]	Reg. CLIP	62.9
SEEN-DA [25]	Reg. CLIP	66.8
DA-Ada [23]	Reg. CLIP	67.3
DA-Faster [5]	FRCNN	38.2
UMT [9]	FRCNN	43.1
PT [4]	FRCNN	55.1
REACT [24]	FRCNN	58.6
DADAOD [16]	FRCNN	69.7
DA-F2F	FRCNN	69.9

Table 4: Ablation study on different combinations of the style-aware feature modulation (SFM) and soft-weighted proposal alignment (SPA) modules.

SFM	SPA	mAP
		57.3
	w/o STA	57.4
	✓	58.1
✓		59.4
✓	✓	60.5

Table 5: Effect of the modulation weight w in Equation (4) on adaptation from Cityscapes to Foggy Cityscapes.

w in Eq. (4)	mAP
0.2	60.0
0.3	60.5
0.4	60.2
0.5	60.0

Cityscapes adaptation scenario. Table 4 reports the performance impact of the style-aware feature modulation (SFM) and the soft-weighted proposal alignment (SPA) modules. For the w/o STA setting, we replace the soft-weighted target alignment (STA) with a conventional label-guided alignment using pseudo-label bounding boxes. As shown in Table 4, the mAP scores exhibit significant degradation when either SFM or SPA is removed. Although applying SPA without STA marginally improves the mAP by 0.1%, the addition of STA achieves significant performance synergy, yielding a further gain of +0.7% mAP. This indicates that SFM effectively transforms source domain representations to align with the target domain via feature-to-feature modulation, while SPA stably separates foreground and background regions for more robust adversarial learning.

Effect of the Modulation Weight w . Table 5 investigates the impact of the modulation weight w within the SFM module. As defined in Eq. (4), w balances the contribution between the original source features and the target-aligned features. The results indicate that DA-F2F maintains robust stability across various configurations of w . Among the tested values, we select $w = 0.3$ as it yields the optimal mAP, striking the best balance for feature modulation.

t-SNE Visualization for SFM. To analyze the modulation impact of the SFM module, we visualize the t-SNE plot of the target features $\mathbf{F}^t$ from Foggy Cityscapes, alongside the source features $\mathbf{F}^s$ and the target-aligned source features $\mathbf{F}^{s \rightarrow t}$ from Cityscapes. Fig. 5(a) displays the t-SNE plot of $\mathbf{F}^t$ and $\mathbf{F}^s$ when



Fig. 4: Qualitative comparison between DADAOD [16] and the proposed DA-F2F across three adaptation scenarios. From the first to fourth columns: input images, detection results of DADAOD and DA-F2F, and the ground-truth annotations.

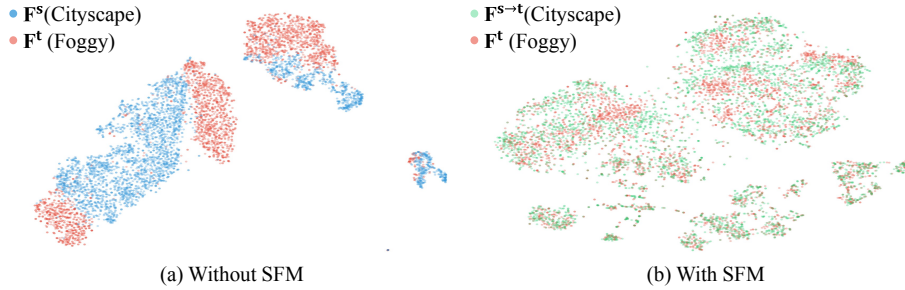


Fig. 5: t-SNE visualization of features from the source (Cityscapes) and target (Foggy Cityscapes) domains. (a) Without SFM, source and target features are clearly separated, indicating a significant domain gap. (b) With SFM, the distributions of the two domains exhibit significantly increased overlap, demonstrating that SFM effectively aligns the domain representations in the feature space.

DA-F2F is trained without the SFM module, clearly illustrating a substantial distribution gap between the two domains. In contrast, Fig. 5(b) demonstrates a significantly increased overlap between the source and target features. This visually confirms that SFM effectively reduces the domain discrepancy at the feature level.

5 Conclusion

In this paper, we proposed DA-F2F, a novel framework for domain adaptive object detection designed to overcome the instability inherent in pixel-level domain transformation. The main components of DA-F2F are the style-aware feature modulation (SFM) and soft-weighted proposal alignment (SPA) modules. SFM effectively transforms source domain representations to align with the target style distribution directly within the feature space. The SPA module advances adversarial learning by stably distinguishing foreground and background regions in the target domain. Comprehensive experiments across various domain adaptation scenarios validate the superiority of DA-F2F.

Acknowledgements

This work was supported by the Institute of Information & Communications Technology Planning & Evaluation(IITP)-Innovative Human Resource Development for Local Intellectualization program grant funded by the Korea government (MSIT) (IITP-2026-RS-2023-00260267).

References

1. Cao, S., Joshi, D., Gui, L.Y., Wang, Y.X.: Contrastive mean teacher for domain adaptive object detectors. In: CVPR. pp. 23839–23848 (2023) [2](#), [3](#), [11](#)
2. Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., Zagoruyko, S.: End-to-end object detection with transformers. In: ECCV. pp. 213–229 (2020) [11](#)
3. Chen, L., Han, J., Wang, Y.: Datr: Unsupervised domain adaptive detection transformer with dataset-level adaptation and prototypical alignment. IEEE TIP **34**, 982–994 (2025) [4](#), [10](#), [11](#), [12](#), [13](#)
4. Chen, M., Chen, W., Yang, S., Song, J., Wang, X., Zhang, L., Yan, Y., Qi, D., Zhuang, Y., Xie, D., Pu, S.: Learning domain adaptive object detection with probabilistic teacher. In: ICML (2022) [2](#), [3](#), [11](#), [12](#), [13](#)
5. Chen, Y., Li, W., Sakaridis, C., Dai, D., Van Gool, L.: Domain adaptive faster r-cnn for object detection in the wild. In: CVPR. pp. 3339–3348 (2018) [3](#), [4](#), [8](#), [11](#), [12](#), [13](#)
6. Chen, Y., Wang, H., Li, W., Sakaridis, C., Dai, D., Van Gool, L.: Scale-aware domain adaptive faster r-cnn. IJCV **129**(7), 2223–2243 (2021) [3](#), [4](#), [8](#)
7. Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., Franke, U., Roth, S., Schiele, B.: The cityscapes dataset for semantic urban scene understanding. In: CVPR. pp. 3213–3223 (2016) [3](#), [10](#)
8. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: CVPR. pp. 248–255 (2009) [10](#)
9. Deng, J., Li, W., Chen, Y., Duan, L.: Unbiased mean teacher for cross-domain object detection. In: CVPR. pp. 4091–4101 (2021) [2](#), [3](#), [4](#), [11](#), [13](#)
10. Deng, J., Xu, D., Li, W., Duan, L.: Harmonious teacher for cross-domain object detection. In: CVPR. pp. 23829–23838 (2023) [2](#), [3](#), [11](#), [12](#), [13](#)
11. DeVries, T., Taylor, G.W.: Improved regularization of convolutional neural networks with cutout. arXiv:1708.04552 (2017) [11](#)
12. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N.: An image is worth 16x16 words: Transformers for image recognition at scale. In: ICLR (2021) [1](#)
13. Gong, K., Li, S., Li, S., Zhang, R., Liu, C.H., Chen, Q.: Improving transferability for domain adaptive detection transformers. In: ACM MM. pp. 1543–1551 (2022) [4](#), [11](#), [12](#), [13](#)
14. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR. pp. 770–778 (2016) [10](#)
15. He, L., Wang, W., Chen, A., Sun, M., Kuo, C.H., Todorovic, S.: Bidirectional alignment for domain adaptive detection with transformers. In: ICCV. pp. 18775–18785 (2023) [4](#), [11](#), [12](#), [13](#)
16. He, X., Li, X., Guo, X.: Differential alignment for domain adaptive object detection. In: AAAI. vol. 39, pp. 17150–17158 (2025) [3](#), [4](#), [8](#), [10](#), [11](#), [12](#), [13](#), [14](#)

17. Hoyer, L., Dai, D., Wang, H., Van Gool, L.: Mic: Masked image consistency for context-enhanced domain adaptation. In: CVPR. pp. 11721–11732 (2023) [11](#)
18. Huang, W.J., Lu, Y.L., Lin, S.Y., Xie, Y., Lin, Y.Y.: Aqt: Adversarial query transformers for domain adaptive object detection. In: IJCAI. pp. 972–979 (2022) [4](#), [11](#), [12](#), [13](#)
19. Huang, X., Belongie, S.: Arbitrary style transfer in real-time with adaptive instance normalization. In: ICCV. pp. 1501–1510 (2017) [7](#)
20. Johnson-Roberson, M., Barto, C., Mehta, R., Sridhar, S.N., Rosaen, K., Vasudevan, R.: Driving in the matrix: Can virtual worlds replace human-generated annotations for real world tasks? arXiv:1610.01983 (2016) [3](#), [10](#)
21. Kennerley, M., Wang, J.G., Veeravalli, B., Tan, R.T.: Cat: Exploiting inter-class dynamics for domain adaptive object detection. In: CVPR. pp. 16541–16550 (2024) [4](#), [11](#), [12](#)
22. Li, H., Zhang, R., Yao, H., Song, X., Hao, Y., Zhao, Y., Chen, Y.: Learning domain-aware detection head with prompt tuning. In: NeurIPS. vol. 36, pp. 4248–4262 (2023) [4](#), [11](#), [13](#)
23. Li, H., Zhang, R., Yao, H., Zhang, X., Hao, Y., Song, X., Li, L.: Da-ada: Learning domain-aware adapter for domain adaptive object detection. In: NeurIPS. vol. 37, pp. 103574–103598 (2024) [4](#), [11](#), [12](#), [13](#)
24. Li, H., Zhang, R., Yao, H., Zhang, X., Hao, Y., Song, X., Li, L.: React: Remainder adaptive compensation for domain adaptive object detection. IEEE TIP **33**, 3735–3748 (2024) [4](#), [10](#), [11](#), [12](#), [13](#)
25. Li, H., Zhang, R., Yao, H., Zhang, X., Hao, Y., Song, X., Li, L.: Seen-da: Semantic entropy guided domain-aware attention for domain adaptive object detection. In: CVPR. pp. 25465–25475 (2025) [4](#), [11](#), [13](#)
26. Li, W., Liu, X., Yuan, Y.: Sigma: Semantic-complete graph matching for domain adaptive object detection. In: CVPR. pp. 5291–5300 (2022) [4](#), [11](#), [12](#), [13](#)
27. Li, Y., Wang, N., Shi, J., Liu, J., Hou, X.: Revisiting batch normalization for practical domain adaptation. arXiv:1603.04779 (2016) [7](#)
28. Li, Y.J., Dai, X., Ma, C.Y., Liu, Y.C., Chen, K., Wu, B., Vajda, P.: Cross-domain adaptive teacher for object detection. In: CVPR. pp. 7581–7590 (2022) [2](#), [3](#), [5](#)
29. Lin, T.Y., Dollár, P., Girshick, R., He, K., Hariharan, B., Belongie, S.: Feature pyramid networks for object detection. In: CVPR. pp. 2117–2125 (2017) [1](#), [10](#)
30. Liu, Y., Wang, J., Huang, C., Wang, Y., Xu, Y.: Cigar: Cross-modality graph reasoning for domain adaptive object detection. In: CVPR. pp. 23776–23786 (2023) [4](#)
31. Liu, Y.C., Ma, C.Y., He, Z., Kuo, C.W., Chen, K., Zhang, P., Wu, B., Kira, Z., Vajda, P.: Unbiased teacher for semi-supervised object detection. arXiv:2102.09480 (2021) [2](#), [5](#)
32. Pan, J., Yin, P., Sun, X., Tan, J., Li, W.: Semi-supervised cross-domain teacher–student mutual training for damaged building detection. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing **16**, 8191–8203 (2023) [2](#), [3](#)
33. Park, T., Efros, A.A., Zhang, R., Zhu, J.Y.: Contrastive learning for unpaired image-to-image translation. In: ECCV. pp. 319–345 (2020) [2](#), [4](#)
34. Perez, E., Strub, F., de Vries, H., Dumoulin, V., Courville, A.C.: Film: Visual reasoning with a general conditioning layer. In: AAAI (2018) [7](#)
35. Redmon, J., Divvala, S., Girshick, R., Farhadi, A.: You only look once: Unified, real-time object detection. In: CVPR. pp. 779–788 (2016) [1](#)
36. Ren, S., He, K., Girshick, R., Sun, J.: Faster r-cnn: Towards real-time object detection with region proposal networks. In: NeurIPS. vol. 28 (2015) [1](#), [9](#), [10](#), [11](#)

37. Sakaridis, C., Dai, D., Van Gool, L.: Semantic foggy scene understanding with synthetic data. *IJCV* **126**(9), 973–992 (2018) [3](#), [10](#)
38. Tarvainen, A., Valpola, H.: Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In: *NeurIPS*. vol. 30 (2017) [2](#), [3](#)
39. Tian, Z., Shen, C., Chen, H., He, T.: Fcos: Fully convolutional one-stage object detection. In: *ICCV*. pp. 9627–9636 (2019) [11](#)
40. Wang, F., Zhao, L., Hong, S., Wang, Z., Liu, C., Gao, C., Luo, D.: Dual-domain teacher for unsupervised domain adaptation detection. *IEEE TMM* **27**, 4217–4226 (2025) [2](#), [3](#), [4](#), [11](#)
41. Wang, W., Cao, Y., Zhang, J., He, F., Zha, Z.J., Wen, Y., Tao, D.: Exploring sequence feature alignment for domain adaptive detection transformers. In: *ACM MM*. pp. 1730–1738 (2021) [4](#)
42. Wang, X., Suo, J., Li, Y., Gao, M., Jiang, P., Dai, P.: Hybrid matching teacher framework for cross-domain visual detection transformer. *IEEE Transactions on Intelligent Transportation Systems* (2025) [3](#), [4](#), [8](#), [10](#), [11](#), [12](#), [13](#)
43. Yoo, J., Chung, I., Kwak, N.: Unsupervised domain adaptation for one-stage object detector using offsets to bounding box. In: *ECCV*. pp. 691–708 (2022) [4](#), [11](#), [13](#)
44. Yu, F., Chen, H., Wang, X., Xian, W., Chen, Y., Liu, F., Darrell, T.: Bdd100k: A diverse driving dataset for heterogeneous multitask learning. In: *CVPR*. pp. 2636–2645 (2020) [3](#), [10](#)
45. Zhang, H., Li, F., Liu, S., Zhang, L., Su, H., Zhu, J., Ni, L.M., Shum, H.Y.: Dino: Detr with improved denoising anchor boxes for end-to-end object detection. In: *ICLR* (2023) [11](#)
46. Zhang, J., Huang, J., Luo, Z., Zhang, G., Zhang, X., Lu, S.: Da-detr: Domain adaptive detection transformer with information fusion. In: *CVPR*. pp. 23787–23798 (2023) [4](#), [11](#), [13](#)
47. Zhang, Y., Wang, Z., Mao, Y.: Rpn prototype alignment for domain adaptive object detector. In: *CVPR* (2021) [3](#), [4](#), [8](#)
48. Zhao, Z., et al.: Masked retraining teacher-student framework for domain adaptive object detection. In: *ICCV* (2023) [2](#), [3](#), [11](#), [12](#), [13](#)
49. Zhong, Y., Yang, J., Zhang, P., Li, C., Codella, N., Li, L.H., Zhou, L., Dai, X., Yuan, L., Li, Y., et al.: Regionclip: Region-based language-image pretraining. In: *CVPR*. pp. 16793–16803 (2022) [11](#)
50. Zhou, H., Jiang, F., Lu, H.: Ssda-yolo: Semi-supervised domain adaptive yolo for cross-domain object detection. *Computer Vision and Image Understanding* **229**, 103649 (2023) [2](#), [3](#)
51. Zhu, J.Y., Park, T., Isola, P., Efros, A.A.: Unpaired image-to-image translation using cycle-consistent adversarial networks. In: *ICCV*. pp. 2223–2232 (2017) [2](#), [4](#)