

Source-Free Object Detection With Detection Transformer

Huizai Yao¹, Sicheng Zhao¹, *Senior Member, IEEE*, Shuo Lu, Hui Chen¹, Yangyang Li, Guoping Liu, *Member, IEEE*, Tengfei Xing, Chenggang Yan¹, Jianhua Tao¹, *Senior Member, IEEE*, and Guiguang Ding¹, *Senior Member, IEEE*

Abstract—Source-Free Object Detection (SFOD) enables knowledge transfer from a source domain to an unsupervised target domain for object detection without access to source data. Most existing SFOD approaches are either confined to conventional object detection (OD) models like Faster R-CNN or designed as general solutions without tailored adaptations for novel OD architectures, especially Detection Transformer (DETR). In this paper, we introduce Feature Reweighting And Contrastive Learning Network (FRANCK), a novel SFOD framework specifically designed to perform query-centric feature enhancement for DETRs. FRANCK comprises four key components: 1) an Objectness Score-based Sample Reweighting (OSSR) module that computes attention-based objectness scores on multi-scale encoder feature maps, reweighting the detection loss to emphasize less-recognized regions; 2) a Contrastive Learning with Matching-based Memory Bank (CMMB) module that integrates multi-level features into memory banks, enhancing class-wise contrastive learning; 3) an Uncertainty-weighted Query-fused Feature Distillation (UQFD) module that improves feature distillation through prediction quality reweighting and query feature fusion; and 4) an improved self-training pipeline with a Dynamic Teacher Updating Interval (DTUI) that optimizes pseudo-label quality. By leveraging these components, FRANCK effectively adapts a source-pre-trained DETR model to a target domain with enhanced robustness and generalization. Extensive experiments on several widely used benchmarks demonstrate that

our method achieves state-of-the-art performance, highlighting its effectiveness and compatibility with DETR-based SFOD models.

Index Terms—Transfer learning, object detection, source-free domain adaptation, contrastive learning.

I. INTRODUCTION

IN THE past decade, deep neural networks [1], [2] have significantly advanced object detection (OD). However, state-of-the-art detectors, such as Faster R-CNN [3] and Detection TRansformer [4] (DETR), require large-scale, high-quality labeled data to achieve optimal performance. Collecting and annotating such data is often expensive and labor-intensive. Furthermore, real-world scenarios frequently exhibit domain shift, where the training data or source domain distribution differs from test data or target domain distribution [5]. This shift severely degrades the generalization ability of conventional OD models, which has led to extensive research on unsupervised domain adaptive object detection (DAOD) [6], [7], [8], [9], [10], [11], [12], [13], [14], [15], [16], [17]. Most DAOD methods rely on adversarial feature alignment [8], [10] or intermediate domain generation [6], [7], both of which require access to annotated source data. However, in real-world applications, source data is often unavailable due to privacy concerns or transmission constraints [18]. In such cases, traditional DAOD techniques become infeasible, necessitating source-free object detection (SFOD) as an alternative.

As illustrated in Fig. 1, SFOD tackles domain adaptation without access to labeled source data. Instead, it relies solely on a source-pre-trained model for adaptation to the target domain [19], [20], [21], [22], [23], [24], [25], [26]. Due to the absence of source data, most SFOD methods employ a Mean Teacher [27] framework with pseudo-labeling to facilitate adaptation. While these approaches have shown promise, they predominantly focus on Faster R-CNN architectures, leveraging components such as RPNs [21], [28]. Consequently, they lack critical insights into adapting DETR-based models [22], [24]. Recent works [23], [29] and a concurrent study [26] have begun exploring SFOD for DETR. However, these efforts either overlook DETR-specific architectural components [23], [29] or focus excessively on teacher-student optimization [26], failing to fully exploit DETR's unique features. Addressing these gaps, we propose a DETR-oriented SFOD framework that effectively incorporates DETR-specific designs for robust adaptation.

Received 18 September 2024; revised 24 March 2025, 21 May 2025, and 20 July 2025; accepted 26 August 2025. Date of publication 16 September 2025; date of current version 19 September 2025. This work was supported in part by Beijing Natural Science Foundation under Grant L252009; in part by the National Natural Science Foundation of China under Grant 62571294, Grant 62441614, and Grant 62525103; in part by Tsinghua University-Fuzhou Joint Institute for Data Technology under Grant JIDT2024012; in part by the Key Research and Development Program of Xinjiang, China, under Grant 2022B01006; and in part by the CCF-DiDi Gaia Collaborative Research Funds. The associate editor coordinating the review of this article and approving it for publication was Dr. Tongliang Liu. (*Corresponding author: Sicheng Zhao.*)

Huizai Yao, Sicheng Zhao, and Hui Chen are with BNRist, Tsinghua University, Beijing 100084, China (e-mail: victoryaohz@gmail.com; schzhao@tsinghua.edu.cn; jichenhui2012@gmail.com).

Shuo Lu is with the Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China (e-mail: lushuo2024@ia.ac.cn).

Yangyang Li is with the Academy of Cyber, Beijing 100846, China (e-mail: liyangyang@ict.ac.cn).

Guoping Liu and Tengfei Xing are with DiDi Chuxing, Beijing 100193, China (e-mail: liuguoping@didiglobal.com; xingtf@foxmail.com).

Chenggang Yan is with the School of Automation, Hangzhou Dianzi University, Hangzhou 310018, China, and also with Lishui Institute of Hangzhou Dianzi University, Lishui 323000, China (e-mail: cgyan@hdu.edu.cn).

Jianhua Tao is with BNRist and the Department of Automation, Tsinghua University, Beijing 100084, China (e-mail: jhtao@tsinghua.edu.cn).

Guiguang Ding is with BNRist and the School of Software, Tsinghua University, Beijing 100084, China (e-mail: dinggg@tsinghua.edu.cn).

Digital Object Identifier 10.1109/TIP.2025.3607621

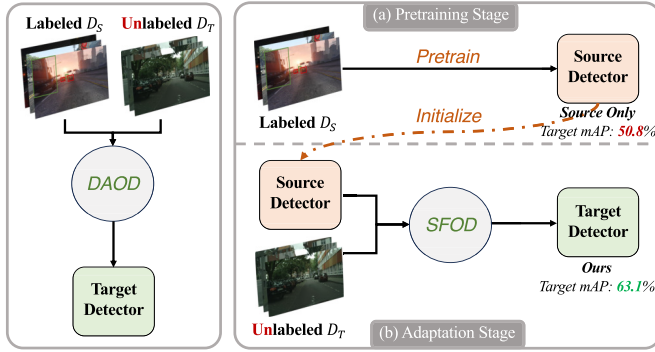


Fig. 1. Illustration of SFOD setting. **Left:** Conventional Domain Adaptive Object Detection (DAOD) approaches utilize both labeled source Domain (D_S) and unlabeled target domain (D_T) to transfer the detector to the target domain. **Right:** Source-Free Object Detection (SFOD) adapts source pre-trained model to the target domain when source data is unavailable.

To address the source-free adaptation challenges faced by DETR, we propose Feature Reweighting AND Contrastive Learning Network (FRANCK), a unified **query-centric** framework that enhances DETR’s adaptation capabilities. We explicitly decompose these challenges into three interconnected levels of alignment: **category-level alignment** (mitigating inter-class confusion), **instance-level alignment** (balancing and supervising samples via pseudo-labels), and **feature-level alignment** (stabilizing cross-domain feature transfer). Guided by this perspective, each module in FRANCK directly targets one level while sharing a common query-centric interface. CMMB enhances category-level alignment by performing class-wise contrastive learning with matching-based memory banks to improve query discriminability. OSSR addresses instance-level alignment by dynamically reweighting query losses through attention-derived objectness scores, mitigating class imbalance and inadequate supervision. UQFD improves feature-level alignment by distilling features with uncertainty-weighted query-fused masks, leading to more stable teacher-student transfer.

These modules form a coherent pipeline and mutually reinforce one another, as illustrated in Fig. 2. Stronger query embeddings from CMMB enable more precise sample weighting in OSSR; enriched queries also guide UQFD to produce more reliable distillation masks; and the stable features obtained from UQFD feed back into both CMMB and OSSR. Through this shared reliance on query representations, FRANCK unifies contrastive learning, sample reweighting, and feature distillation across category-level, instance-level, and feature-level alignment, resulting in synergistic improvements in both discriminability and transferability for DETR-based SFOD.

Our main contributions are as follows:

- We systematically explore the challenges of SFOD on DETRs, an area that has received limited attention, and propose a novel framework that explicitly incorporates DETR-specific architectural designs.
- We propose FRANCK, a novel framework that introduces several key innovations tailored for DETR-based source-free domain adaptive object detection. Motivated by a

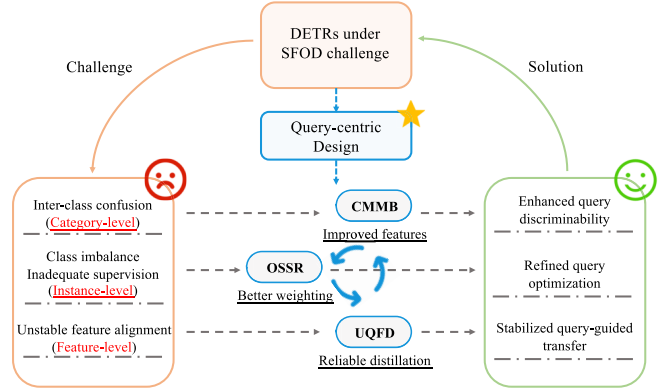


Fig. 2. A conceptual framework illustrating how FRANCK addresses DETR’s source-free challenges through a unified query-centric design. The three challenges are organized as category-level alignment (inter-class confusion), instance-level alignment (class imbalance and inadequate supervision), and feature-level alignment (unstable feature alignment). Each module (CMMB, OSSR, UQFD) targets one of these challenges, while their shared reliance on query representations forms a synergistic loop where improved features, better weighting, and reliable distillation reinforce each other for robust and efficient adaptation.

query-centric representation enhancement principle, the components cooperate to effectively adapt the model to the target domain and improve detection performance.

- We conduct extensive experiments on several widely-used benchmarks, demonstrating that FRANCK achieves state-of-the-art performance in SFOD for DETRs.

The remainder of this paper is structured as follows: Section II provides a comprehensive review of related work, covering Object Detection (OD), Domain Adaptive Object Detection (DAOD), Source-Free Domain Adaptation (SFDA), and Source-Free Object Detection (SFOD). Section III details the proposed FRANCK framework and its key components. Section IV presents experimental results, including quantitative analysis, ablation studies, and visualization experiments, along with essential implementation details. Finally, Section V summarizes our findings and concludes the paper.

II. RELATED WORK

This section provides an overview of research relevant to our method, covering OD, DAOD, SFDA, and SFOD.

A. Object Detection

Object detection (OD) seeks to identify and localize objects in images. With the rise of deep learning and convolutional neural networks (CNNs) [2], OD has made significant progress. Traditional methods are typically categorized into two-stage detectors (e.g., R-CNN [30], Faster R-CNN [3]) that generate region proposals before classification, and one-stage detectors (e.g., SSD [31], FCOS [32], YOLO [33]) that predict objects directly. While efficient, these models often rely on heuristic components like Non-Maximum Suppression (NMS), making them sensitive to hyperparameters.

Transformer-based models have redefined OD by formulating it as a set prediction problem. DETR (DEtection TRansformer) and its variants [4], [34], [35], [36], [37], [38],

[39] remove the need for NMS, enabling end-to-end detection. Variants such as Deformable DETR [38] and DN-DETR [37] improve convergence and robustness. However, DETR-based models often underperform under domain shifts and have limited cross-domain generalization [12], [13], [14], [15]. In particular, DETR-based source-free object detection remains underexplored. To bridge this gap, we adopt DETR as our base and investigate its extension to source-free cross-domain object detection.

B. Domain Adaptive Object Detection

Domain Adaptive Object Detection (DAOD) aims to mitigate domain shifts and enhance the generalization ability of object detectors. Most DAOD research has focused on traditional detection frameworks such as Faster R-CNN [40], [41], [42], [43], [44], [45], [46], [47], [48], FCOS [8], [9], [10], [11], and YOLO [49], [50]. Recently, DAOD methods tailored for DETRs have emerged. For example, SFA [12] aligns features at both encoder and decoder levels, while MTTrans [13] adopts multi-level feature alignment within a Mean Teacher [27] framework. Other approaches, such as DA-DETR [14], integrate CTBlender with Split-Merge Fusion and Scale Aggregation Fusion for better alignment. MTM [15] utilizes mask-integrated adversarial alignment and mixed queries to ensure consistent learning, while BiADT [16] introduces bi-directional domain alignment with token-wise domain embeddings. Similarly, ACCT [17] employs adversarial alignment, confidence thresholding, and contrastive learning to tackle domain shifts. Despite their effectiveness, these methods rely on source data and labels, making them inapplicable in scenarios where source data is unavailable. Our work addresses this limitation by enabling DETR-based DAOD in a source-free setting.

C. Source-Free Domain Adaptation

Conventional Unsupervised Domain Adaptation (UDA) relies on source domain data and labels, limiting its applicability in privacy-sensitive scenarios [18], [51] and under data transmission constraints [25], [52]. To address this, Source-Free Domain Adaptation (SFDA) has been introduced, enabling adaptation using only unsupervised target data and a pre-trained source model, without direct access to source data. SFDA has been successfully applied to tasks like image classification [53], [54], semantic segmentation [55], human pose estimation [56], gesture recognition [57], and panoramic segmentation [58]. It leverages various model fine-tuning strategies, with semi-supervised knowledge distillation in a teacher-student framework [52] being a widely adopted approach. Additionally, contrastive learning has proven effective in improving generalization by learning discriminative feature representations [59], [60].

Despite SFDA's success in classification and segmentation, its direct application to object detection is challenging. Unlike these tasks, object detection requires both classification and localization of multiple objects, demanding specialized architectures like Faster R-CNN [3] and DETR [4]. To tackle this, Source-Free Object Detection (SFOD) has emerged, enabling

domain adaptation for detection tasks. Our work extends SFOD to DETR-based detectors, an area largely unexplored, enhancing DETR's adaptability under domain shifts.

D. Source-Free Object Detection

To tackle the challenge of source-free adaptation in object detection, researchers have developed Source-Free Object Detection (SFOD) methods. SED [19] improves detection performance using self-entropy descent and mosaic augmentation [61], while LODS [21] employs style transfer modules and multi-level feature alignment to minimize domain discrepancy. IRG-SFDA [20] constructs information relation graphs to enhance knowledge distillation and contrastive learning. AASFOD [22] applies adversarial alignment on target self-divided data obtained via Monte-Carlo sampling. Meanwhile, Balanced Teacher (BT) [41] introduces class-balanced instance selection and progressive target variance minimization to mitigate imbalance issues. DACA [25] extends SFOD to multi-source scenarios by incorporating region proposal fusion, pseudo-label ensembling, and class-wise contrastive learning. However, most of these methods either depend on Faster R-CNN-specific components, such as region proposal networks (RPNs), or fail to consider modern DETR architectures.

Notably, a few studies have explored the feasibility of SFOD on DETRs. TeST [23] develops a two-stage self-training process that adapts the teacher and student networks separately, but lacks specific designs tailored to DETR components that could further enhance performance. A concurrent study, DRU [26], employs masked image consistency [62] and dynamic retraining-updating [63] for effective SFOD on DETRs. Nevertheless, DRU [26] focuses more on general self-training and updating mechanisms, overlooking effective DETR feature adaptation that could further improve SFOD performance. In contrast, our proposed method is carefully designed for the DETR architecture, enabling efficient feature learning and improving detection performance on DETRs.

III. PROPOSED METHOD

A. Preliminaries

1) *Problem Setup*: We first introduce the problem setup for SFOD tasks. Unless otherwise specified, we consider SFOD tasks under an unsupervised domain adaptation (UDA) setting, where the target labels are entirely unavailable. In the SFOD setting, there exists a source domain D_S sampled from source distribution $p_S(x_S, y_S)$ and a target domain D_T sampled from target distribution $p_T(x_T, y_T)$, where x denotes image and y denotes corresponding label. We follow the closed-set DA setting, in which D_S and D_T both have k foreground categories. Since source data and distribution are unavailable for adaptation, only a source pre-trained model θ_S and unsupervised target dataset $X_T = \{x_T^i\}_{i=1}^{N_T}$ are available. Our goal is to perform effective adaptation, obtaining a detection model $\theta_T : x_T \rightarrow y_p$ that works well on the target domain. We utilize and focus on DETR [4], [38] structure for detection. We use Deformable DETR [38] as the base detector and denote the total number of object queries in one DETR model as n_q .

2) *Mean-Teacher-Based SFOD*: In many SFDA and SFOD applications, the Mean Teacher (MT) framework [27] is a key self-training approach that enables model adaptation without target-domain supervision [13], [20], [21], [22], [25], [26]. Originally designed for semi-supervised learning, MT leverages strong-weak augmentation and consistency regularization for knowledge transfer. A key feature is its Exponential Moving Average (EMA) update, ensuring stable parameter update.

In SFOD, both the teacher and student models are initialized with an identical source-pre-trained network. During training, a target-domain sample from D_T undergoes strong and weak augmentations, which are then fed into the student and teacher networks, respectively. Without supervision, the teacher generates pseudo-labels by confidence-thresholding weakly augmented sample predictions. The student model then updates its parameters by minimizing the following loss:

$$\mathcal{L}_{\text{det}} = \mathcal{L}_{\text{cls}} + \mathcal{L}_{\text{reg}} + \mathcal{L}_{\text{aux}}, \quad (1)$$

where $\mathcal{L}_{\text{det}}$ is the detection loss of DETRs, $\mathcal{L}_{\text{cls}}$ and $\mathcal{L}_{\text{reg}}$ the classification and regression loss, respectively. $\mathcal{L}_{\text{aux}}$ is the auxiliary loss, if applicable. Following Deformable DETR [38], we adopt Focal Loss [64] as the classification loss. The student parameters Θ_{stu} are updated via backpropagation, while the teacher parameters Θ_{tea} follow EMA updates:

$$\left\{ \begin{array}{l} \Theta_{\text{stu}} \leftarrow \Theta_{\text{stu}} + \eta \frac{\partial(\mathcal{L}_{\text{stu}})}{\partial \Theta_{\text{stu}}}, \\ \Theta_{\text{tea}} \leftarrow \alpha_{\text{EMA}} \Theta_{\text{tea}} + (1 - \alpha_{\text{EMA}}) \Theta_{\text{stu}}, \end{array} \right. \quad (2a)$$

$$\quad (2b)$$

where η is the student learning rate and α_{EMA} the EMA update rate. MT provides robust model optimization and adaptation in semi-supervised and cross-domain settings, making it a common baseline for many DAOD and SFOD approaches.

B. Overview

In this section, we outline our problem setup, the Mean-Teacher-based SFOD architecture with its update mechanism, and the key components of FRANCK. CMMB leverages pseudo-label-induced bipartite matching to build class-wise memory banks for contrastive learning, enhancing feature discriminability. OSSR mitigates class imbalance by assigning dynamic instance-wise loss weights via a query-fused objectness score. UQFD improves knowledge transfer through uncertainty-weighted, objectness-guided feature distillation. Finally, we present the overall training loss and DTUI, which strengthens Mean Teacher robustness by dynamically adjusting the EMA update interval.

C. Contrastive Learning With Matching-Based Memory Bank

While the original MT framework with pseudo-labeling lays a solid foundation for SFOD performance, feature representation remains suboptimal [20], [25]. To address this, we adopt class-wise contrastive learning following prior studies [25], [45], [65] and fuse multi-level decoder query features to enhance learning. Additionally, given the class imbalance problem in OD tasks and inspired by bipartite matching in DETRs, we introduce memory banks and a pseudo bipartite matching strategy for class-wise contrastive learning.

1) *Class-Wise Contrastive Learning*: Contrastive learning enhances model discriminability by pulling positive samples closer and pushing negatives apart. We build our contrastive loss on the Supervised Contrastive Loss (SCL) [65]. Suppose we have c sample groups, denoted as $\mathcal{K}_a = \mathcal{K}_0 \cup \mathcal{K}_1 \cup \dots \cup \mathcal{K}_{c-1}$ where each group corresponds to a category. Following SCL, the class-wise contrastive loss is formulated as:

$$\mathcal{L}_{\text{cont}} = \frac{1}{c} \sum_{i=0}^{c-1} \frac{-1}{|\mathcal{K}_i|} \sum_{Q \in \mathcal{K}_i} \sum_{K^+ \in \mathcal{K}_i} \log \frac{\exp(Q \cdot K^+ / \tau)}{\sum_{K \in \mathcal{K}_a} \exp(Q \cdot K / \tau)}, \quad (3)$$

where Q is the *contrastive learning query feature* that attracts positive keys and repels negative keys in contrastive learning.

2) *Query Feature Fusion and Memory Bank*: For contrastive learning in object detection, it's intuitive to use class-wise instance-level features directly as contrastive samples. However, this can be ineffective or even detrimental because (1) unlike Faster R-CNN, which generates multiple contrastive samples per object via the anchor mechanism [3], [20], DETR aims to assign only one query per object at a time [4]; and (2) real-world class distributions are often imbalanced, leading to significant biases. To mitigate these issues, we adopt a simple yet effective memory bank technique [25], [66], [67]. We construct $k+1$ memory banks $\{\mathcal{M}_i\}_{i=0}^{k+1}$, one for each category, including a background memory bank $\mathcal{M}_0$. Background features are considered because DETR assigns queries to different objects and naturally generates diverse negative samples, making background features valuable for contrastive learning. Each memory bank maintains a fixed maximum size $l_{\mathcal{M}}$ and is updated using a First-In-First-Out (FIFO) strategy. The non-empty memory banks are denoted as $\{\mathcal{M}_i\}_{i=0}^{v+1}$, where $\mathcal{M}_0$ corresponds to the background memory bank and $v \leq k$. The contrastive loss with memory banks is:

$$\mathcal{L}_{\text{cont}} = \frac{1}{v} \sum_{i=1}^{v+1} \frac{-1}{|\mathcal{M}_i|} \sum_{Q \in \mathcal{M}_i} \sum_{K^+ \in \mathcal{M}_i} \log \frac{\exp(Q \cdot K^+ / \tau)}{\sum_{K \in \mathcal{M}_a} \exp(Q \cdot K / \tau)}, \quad (4)$$

where $\mathcal{M}_a$ represents all samples across memory banks. In Eq. (4), the class index i starts from 1, as background features are only considered negative samples. To further exploit the semantic information contained in multi-scale transformer features, we fuse multi-scale decoder output features of object queries via scale-wise summation. This strategy enhances the representation of instance-level features, providing richer contextual information for improved contrastive learning.

3) *Pseudo Bipartite Matching Assignment*: In the context of contrastive learning for SFOD, a significant challenge stems from the lack of ground truth objects in the target domain, which complicates the construction of contrastive pairs. Due to the set-prediction mechanism of DETRs, RoI-based region features extracted from the backbone [20], [25], [45] (e.g., using RoIAlign [68]) are decoupled from the instance representations formed by the transformer and object queries. Thus, contrastive learning applied only at the backbone level cannot directly enhance the discriminative power of the object queries, making contrastive learning on object queries a more suitable choice for DETR-based methods. To perform contrastive learning on object queries, a natural approach

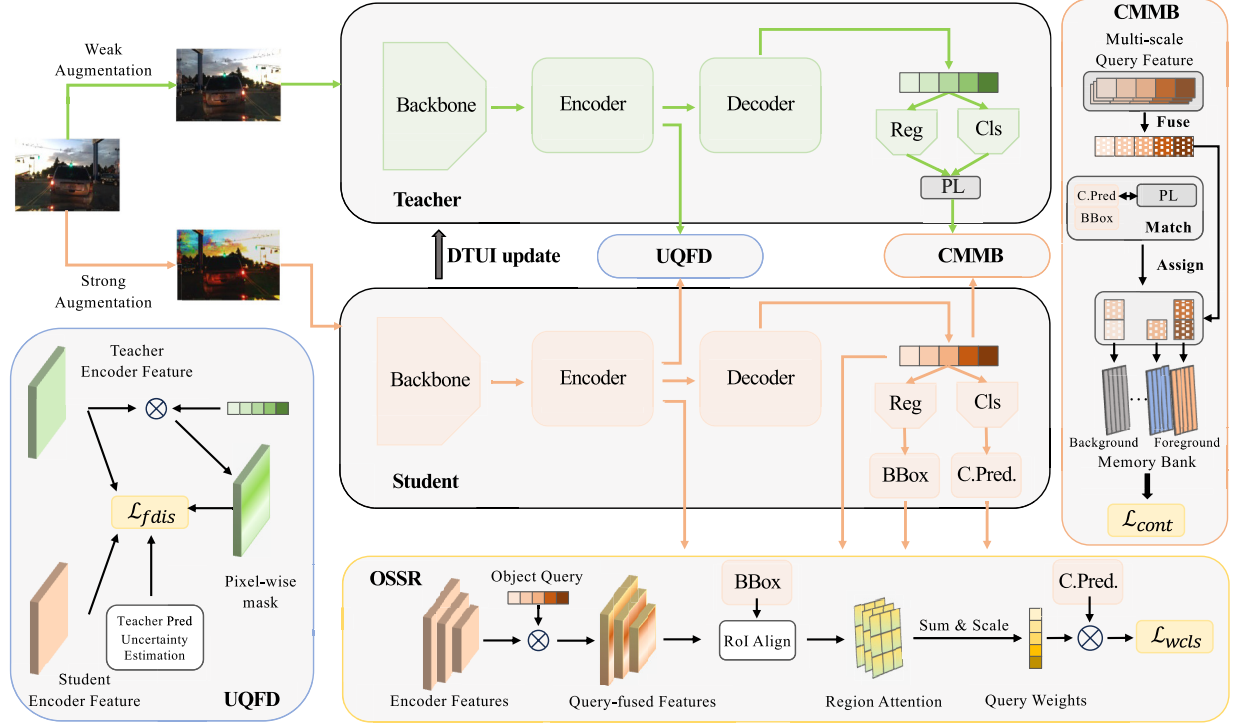


Fig. 3. The proposed Feature Reweighting AND Contrastive Learning Network (FRANCK). Source data is only available at the source pretraining stage. Within FRANCK, the teacher and student models collaborate to optimize the student network through UQFD, OSSR, and CMMB. The teacher network is updated dynamically using DTUI, ensuring more stable and effective adaptation. “Reg” and “Cls” refer to the regression and classification heads of DETR, respectively. “BBox” and “C.Pred.” denote the bounding box predictions and classification predictions, respectively, while “PL” denotes pseudo-labels.

is to select foreground features from all n_q query features under DETRs. A straightforward yet suboptimal method is to filter queries based on prediction confidence. Specifically, if a query’s highest probability corresponds to a foreground class and surpasses a predefined threshold, it is considered a foreground feature; otherwise, it is classified as background. However, this approach is heuristic and lacks robustness.

Instead, we leverage *bipartite matching*, one of the core assignment mechanisms in DETRs [4]. Since DETRs assign queries to ground-truth labels via bipartite matching, we extend this principle to contrastive pair construction. More specifically, queries matched with pseudo-labels corresponding to an i -th foreground class are directly taken as foreground features, while unmatched queries are classified as background. This assignment strategy naturally integrates with DETRs and provides a more stable and unbiased feature selection mechanism. To this end, we adopt bipartite-matching-based contrastive pair assignment between student query features and pseudo-labels as our final strategy for constructing contrastive pairs, as illustrated in Fig. 3. By leveraging pseudo bipartite matching along with the memory bank, we achieve robust class-wise contrastive learning in SFOD for DETRs.

D. Objectness Score Sample Reweighting

Class imbalance is a common challenge in object detection tasks. A common solution is soft sampling, *i.e.*, reweighting based on prediction quality [64], [69] or IoU quality [70]. However, in SFOD tasks, the absence of ground-truth labels makes it difficult to estimate prediction quality accurately.

To overcome this, inspired by [71], [72] and [73], we adopt the *objectness score*, which leverages intrinsic feature properties of feature maps. Following these approaches, we first extract and normalize multi-scale encoder features $\{\mathcal{F}_e^i | i = 3, 4, 5\}$ and decoder query features $\mathcal{F}_q$, where $\mathcal{F}_e^i \in \mathbb{R}^{H_i \times W_i \times C}$ and $\mathcal{F}_q \in \mathbb{R}^{n_q \times C}$, together with query bounding boxes as $bbox \in \mathbb{R}^{n_q \times 4}$. To focus encoder features on query-relevant object information, we perform query-wise fusion and summation to obtain query-fused encoder features $\{\mathcal{F}_{eq}^i | i = 3, 4, 5\}$:

$$\mathcal{F}_{eq}^i = \frac{1}{n_q} \sum_{j=0}^{n_q-1} \mathcal{F}_e^i \cdot \mathcal{F}_q^T[j], \quad (5)$$

where $\mathcal{F}_{eq}^i \in \mathbb{R}^{H_i \times W_i}$ represents the fused encoder feature across all n_q queries. Note that since the teacher model in the Mean Teacher (MT) framework undergoes a more stable updating process and encodes more robust features compared student, we leverage $\mathcal{F}_e^i$ from the teacher encoder to ensure stable and reliable knowledge transfer. Once the attention map is obtained, prior works [71], [73] typically upsample the feature map to the original image scale and pooled bounding box features using approximate integer coordinates. However, this naïve approach can lead to information loss and inadequate extraction of small object features, introducing significant bias. Instead, we adopt RoIAlign [68], which uses bilinear interpolation [74] for precise feature extraction and information loss mitigation. Note that this use of RoIAlign does not contradict our earlier statement regarding its limitations in CMMB. Here, RoIAlign is adopted to improve the precision of attention

score pooling, particularly in the case of small objects, not for enhancing the object detection model through contrastive learning. It thus acts as a natural and effective substitute for direct pooling. Specifically, we apply RoIAlign to extract proposal features $\mathcal{F}_{\text{align}}^i$ from the corresponding query-fused encoder features $\mathcal{F}_{\text{eq}}^i$:

$$\mathcal{F}_{\text{align}}^i = \text{RoIAlign}(\mathcal{F}_{\text{eq}}^i, \text{bbox}), \quad (6)$$

where $\mathcal{F}_{\text{align}}^i \in \mathbb{R}^{n_q \times H_a \times W_a}$, with H_a and W_a representing the height and width of the RoIAlign output, respectively. Each query is then assigned an attention score of $H_a \times W_a$, enabling us to compute a set of objectness scores:

$$\mathcal{S} = \left\{ \sum_{i=3,4,5} \sum_{H_a \times W_a} \mathcal{F}_{\text{align}}^i[j] \right\}_{j=0}^{n_q-1}, \quad (7)$$

The computed objectness scores $\mathcal{S}$ encode query-wise attention over multi-scale features. We observe that utilizing all object queries often leads to higher objectness attention for easily detectable foreground or background objects, aligning with findings in DETRDistill [72]. To mitigate foreground-background class imbalance and improve model discrimination for hard samples, we assign higher weights to foreground areas and hard samples with low attention scores. Specifically, we employ these normalized weights to refine the original Focal Loss [64] in Deformable DETR [38] for classification:

$$\mathcal{L}_{\text{wcls}} = \frac{1}{n_q} \sum_{i=0}^{n_q-1} w_i [-\alpha_t (1 - p_i)^\gamma \log(p_i)], \quad (8)$$

and

$$w_i = (1 - \text{MinMaxScaler}(\mathcal{S}[i]))^\beta, \quad (9)$$

where $\text{MinMaxScaler}(\mathcal{S}[i]) = \frac{\mathcal{S}[i] - \min(\mathcal{S})}{\max(\mathcal{S}) - \min(\mathcal{S})}$, and β is a smoothing hyperparameter. α_t and γ are the Focal Loss balancing parameters. By reweighting the query loss with attention-based objectness scores, our method encourages the model to focus more on foreground regions and hard samples. This enhances mutual learning under the teacher model's pseudo supervision and improves model discriminability.

E. Uncertainty-Weighted Query-Fused Feature Distillation

To ensure robust distillation between the teacher and student networks in SFOD, consistency regularization [20], [25], [75] is a widely adopted approach. However, existing methods are primarily designed for Faster R-CNN, where shared proposals naturally align classification and localization scores for consistency regularization. In contrast, DETR employs a query-based mechanism, where different queries correspond to different objects, making direct consistency loss computation (e.g., the KL divergence) based on prediction indices infeasible. For instance, in most cases, the i -th predictions from the teacher and student networks correspond to different objects, rendering direct consistency enforcement impractical.

To address this challenge, we opt for feature imitation instead of logit mimicking for knowledge distillation, a strategy that has proven effective in object detection tasks [72], [76], [77]. Given that the student network in the original Mean

Teacher framework processes strongly augmented images, we introduce an additional forward pass of these augmented images through the teacher model. This forward pass is solely used to extract image-level features for distillation. Following DETRDistill [72], we reweight query-fused features based on prediction quality and formulate a unified feature distillation loss, denoted as $\mathcal{L}_{\text{fdis}}$.

1) *Uncertainty-Based Query Weighting*: We begin by describing our approach to extracting features for distillation. Utilizing object queries and image-level features, we construct objectness-weighted feature maps, similar to OSSR. However, DETRDistill [72] found that naïve objectness-weighted feature distillation is ineffective due to the varying contributions of queries. To address this, DETRDistill applies soft attention masks with quality scores [78] derived from GT labels and teacher predictions. However, this does not apply to SFOD tasks, where GT labels are unavailable. To overcome this limitation, we leverage uncertainty estimation, using prediction entropy as the quality score. Specifically, we extract prediction scores from queries and compute entropy for teacher model predictions, denoted as $\mathcal{E} \in \mathbb{R}^{n_q}$. To assign higher weights to more reliable regions identified by the robust teacher model, we normalize the query weights using $\mathcal{W}_q = (1 - \text{MinMaxScaler}(\mathcal{E}))^{\beta'}$, where we set $\beta' = 1$ for simplicity.

2) *Query-Fused Feature Distillation*: Following DETRDistill, we compute feature distillation loss by applying soft masks across queries and derive a unified weighted loss. Given a single query feature $\mathcal{F}_q^j$ and the last encoder layer feature $\mathcal{F}_e$, we derive the query-fused feature $\mathcal{F}_{\text{eq}}^j \in \mathbb{R}^{H \times W}$ by:

$$\mathcal{F}_{\text{eq}}^j = \mathcal{F}_e \cdot (\mathcal{F}_q^j)^T. \quad (10)$$

Then by using $\mathcal{F}_{\text{eq}}^j$ as an objectness-based soft mask, we can perform weighted feature imitation by:

$$\mathcal{L}_{\text{fdis}} = \frac{1}{n_q HWC} \sum_{j=0}^{n_q-1} \mathcal{W}_q^j \|\mathcal{F}_{\text{eq}}^j \odot (\mathcal{F}_T - \mathcal{F}_S)\|_2^2, \quad (11)$$

where $\mathcal{W}_q^j$ and $\mathcal{F}_{\text{eq}}^j$ denote the j -th elements of $\mathcal{W}_q$ and the query-fused feature, respectively. $\mathcal{F}_T$ and $\mathcal{F}_S$ represent the last encoder layer features of the teacher and student models, respectively, and $\odot$ denotes Hadamard product. By incorporating weighted feature distillation, the teacher guides the student to generate more stable feature representations, improving the robustness of knowledge transfer.

F. Dynamic Teacher Updating Interval

To enhance the conventional Mean Teacher self-training approach, we introduce an improved updating mechanism termed Dynamic Teacher Updating Interval (DTUI). As discussed in Sec. III-A, we adopt the Mean Teacher framework to ensure a robust and efficient adaptation scheme. Specifically, after performing forward propagation on both the teacher and student networks, we first filter pseudo-labels by applying a confidence threshold of 0.3 to the teacher's predictions. We then compute the total loss function as follows:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{wcls}} + \mathcal{L}_{\text{reg}} + \mathcal{L}_{\text{aux}} + \omega_1 \mathcal{L}_{\text{cont}} + \omega_2 \mathcal{L}_{\text{fdis}}, \quad (12)$$

where $\mathcal{L}_{wcls}$, $\mathcal{L}_{reg}$, and $\mathcal{L}_{aux}$ collectively form the detection loss $\mathcal{L}_{det}$ by replacing the original classification loss with Eq. (8). The student model is then optimized using $\mathcal{L}_{total}$, while the teacher model is updated using EMA with a momentum factor of α_{EMA} .

1) *DTUI*: In SFOD tasks, domain shifts and the absence of ground truth labels can destabilize the student model's prediction and optimization, leading to biased teacher updates and reduced mutual learning effectiveness [25], [63]. To mitigate this issue, AASFOD [22] and DACA [25] adopt fixed EMA update intervals for each experiment, which is helpful but ignores the adaptation progress over time. In contrast, we propose a dynamic EMA interval i_{EMA} , formulated as:

$$i_{EMA} = \delta + \lfloor e/\epsilon \rfloor, \quad (13)$$

where δ is a base interval that controls the steps for stable knowledge accumulation, e represents the current epoch index, and ϵ denotes the increment rate. Under this dynamic strategy, while the student network is updated at every iteration using Eq. (2a), the teacher network is updated every i_{EMA} iterations. This linear interval adjustment allows for frequent parameter exploration at the early stages of adaptation, facilitating a more effective search of the parameter space, and progressively stabilizing the model's updates as training progresses.

IV. EXPERIMENTS

This section presents the datasets, experimental setup, experimental results, and comprehensive analyses from quantitative, ablation, and visualization studies to validate our method.

A. Datasets

We evaluate our approach on four widely used object detection datasets: Cityscapes [81], Foggy Cityscapes [82], Sim10k [83], and BDD100K [84], along with a synthesized rainy Cityscapes dataset. Our experiments cover cross-weather, synthetic-to-real, and cross-scene adaptation.

1) *Cross-Weather Adaptation*: Cityscapes [81] is an urban scene dataset with 2,975 training and 500 validation images from various cities. Foggy Cityscapes [82] extends it via synthetic fog generation. We use Cityscapes and its foggy variant (with 0.02 fog density) as the source and target domains, respectively. To further evaluate robustness under adverse weather, we introduce a rainy version of Cityscapes using RainMix [85], following prior works [79], [86], enabling assessment across diverse weather conditions [79], [87], [88].

2) *Synthetic-to-Real Adaptation*: Sim10k [83], generated from the GTA V game, contains 9,000 training and 1,000 validation images. In this setting, Sim10k is the source domain and Cityscapes is the target domain. This setting assesses the ability to generalize from synthetic data to real data distributions, offering benefits such as reduced data collection costs and enhanced data diversity in real-world scenarios.

3) *Cross-Scene Adaptation*: BDD100K [84] is a large-scale autonomous driving dataset covering different times of the day. Following prior work [12], [13], [26], we use only daytime images, comprising 36,728 training and 5,258

validation images. In this setting, Cityscapes is used as the source domain and BDD100K daytime as the target domain. This assesses a detection model's adaptability across diverse scenes.

4) *Cross-Dataset Adaptation*: KITTI [89] is an autonomous driving dataset collected from diverse real-world scenes. In our setting, all 7,481 annotated images from KITTI are used as the source domain, while Cityscapes serves as the target domain. This setup evaluates the detection model's ability to adapt across different camera systems and dataset characteristics.

5) *Cityscapes-to-ACDC Adaptation*: ACDC [90] is a dataset designed for comprehensive understanding of autonomous driving scenes. It encompasses four types of challenging real-world weather conditions, including snow, rain, night, and fog. This adaptation setting is used to further evaluate the effectiveness and robustness of our method under diverse and complex real-world domain shifts.

B. Baselines

We compare our method against multiple baseline settings, including source-only, DAOD, SFOD, and Oracle.

1) *Source-only*: For Source-only baselines, the source pre-trained model is directly evaluated on the target domain without adaptation, serving as a lower bound for domain adaptation.

2) *DAOD*: We compare our approach with previous DETR-based DAOD methods, including SFA [12], MTTrans [13], DA-DETR [14], MTM [15], BiADT [16], and ACCT [17]. Among them, only BiADT uses DAB-Deformable-DETR [36], a variant of Deformable DETR [38], as its base detector, while all others adopt Deformable DETR. Additionally, we also compare with approaches based on different detectors, including SIGMA++ [8], CIGAR [9], EPM [10], and MGA-DA [11]. These comparisons provide insights into DAOD performance on DETRs and highlight our method's advantages.

3) *SFOD*: We include DETR-based SFOD methods such as TeST [23] and the concurrent work DRU [26]. Since TeST is not open-source, we implement and reproduce it, tuning hyperparameters for optimal results. OnDA-DETR [29] is omitted due to its similarity to vanilla Mean Teacher, for which we conduct separate experiments. Given the limited number of DETR-based SFOD works, we also compare against Faster R-CNN-based SFOD methods, including SED [19], IRG-SFDA [20], AASFOD [22], and BT [24].

4) *Oracle*: For Oracle baselines, models are trained and tested directly on the labeled target domain without source pretrain, representing the upper bound of DAOD to some extent.

Following prior work, detection performance is evaluated using mean average precision (mAP) with IoU=0.5. From Tab. II to Tab. IV, we use R-50 and R-101 to refer to ResNet-50 and ResNet-101 [2], respectively, and V-16 to denote the VGG-16 [93] network. Our comparisons include methods based on FCOS [32], DETR [4], [38], and Faster R-CNN (FRCNN) [3]. Notably, apart from BiADT [16], which adopts DAB-Deformable DETR [36], which is a variant of Deformable DETR [38], all DETR-based approaches in our study are built upon Deformable DETR [38].

TABLE I
HYPERPARAMETERS SETTING

Hyperparam	Explanation	Value
l_M	Maximum length of a memory bank	100
H_a, W_a	Height and width of RoI Align output	7, 7
α_t, γ	Balancing parameter of Focal Loss	0.25, 2
β	Smoothing parameter for $\mathcal{L}_{wcls}$	0.2
β'	Smoothing parameter for $\mathcal{W}_q$ in $\mathcal{L}_{fdis}$	1
τ	Smoothing parameter for $\mathcal{L}_{cont}$	0.07
α_{EMA}	EMA updating rate	0.999
ω_1	Weighting factor of $\mathcal{L}_{cont}$	0.4
ω_2	Weighting factor of $\mathcal{L}_{fdis}$	0.1
δ	DTUI base interval	5
δ	DTUI base interval (Cross-Scene Adaptation)	60
ϵ	DTUI interval increasing rate	5
c_{thresh}	Confidence threshold for pseudo-labeling	0.3

C. Implementation Details

In this section, we provide the implementation details of the experiments. The hyperparameters, along with their notations, descriptions, and values, are summarized in Tab. I. In the source pretraining stage, we train the model for 50 epochs, starting with a learning rate of 2×10^{-4} , which is reduced by a factor of 0.1 after 40 epochs. During the adaptation stage, the model is trained for 30 epochs with a fixed learning rate of 5×10^{-5} . All experiments use a batch size of 2 per GPU, with training conducted on 4 NVIDIA RTX 4090 GPUs.

D. Comparison With State-of-the-Art

In this section, we present quantitative comparisons with state-of-the-art methods. The results are summarized in Tab. II, Tab. III, Tab. IV, Tab. V, and Tab. VI, with the best mAP(%) values and second-best mAP values (excluding Oracle results) bolded and underlined, respectively. Our method achieves state-of-the-art results across multiple domain adaptation settings. In cross-weather adaptation, it attains 44.9 mAP on Cityscapes to Foggy Cityscapes, showing that the model remains robust and effectively overcomes domain shift even when object visibility is severely reduced, and reaches 50 mAP for rainy conditions. In synthetic-to-real adaptation, it achieves 63.1 mAP, demonstrating strong transferability from cost-effective synthetic data to complex real scenes. In Cityscapes-to-ACDC adaptation, it surpasses DRU across all weather conditions with gains of 1.0 to 4.6 mAP, confirming its ability to handle diverse environments. In cross-scene and cross-dataset adaptation, it reaches 40.7 mAP and 48.5 mAP respectively, further proving its robustness across varying locations and data sources.

E. Ablation Study

In this section, we present several ablation studies to assess the design and effectiveness of our method. Unless specified, we conduct all ablation studies on Cityscapes to Foggy Cityscapes adaptation in the Cross-weather setting.

1) *Component Ablation*: We evaluate the impact of different components by comparing performance with and without each component. As shown in Tab. IX, the ablation results

demonstrate that (1) the Mean Teacher strategy and dynamic MT updating interval improve SFOD performance by enhancing robust training and knowledge distillation, and (2) The proposed methods, including CMMB, OSSR, and UQFD, each contribute to the final detection performance, resulting in a 3.2 mAP gain based on Mean Teacher with DTUI with only 28.5% additional training time.

2) *Effectiveness Across Backbones*: While our main experiments use ResNet-50, we further evaluate transformer-based backbones, including ViT-Base [94] and Swin Transformers [95] of different scales—Swin-T (tiny), Swin-S (small), Swin-B (base), and Swin-L (large). As shown in Tab. VII, vanilla ViTs perform worse overall due to their single-scale, low-resolution features [96], yet the trends remain consistent: (1) domain shifts lead to similar performance degradation across all backbones, and SFOD methods effectively mitigate this issue; and (2) our method consistently surpasses MT [27] and DRU [26] on every backbone tested, highlighting both its robustness and its ability to generalize across different feature extractors.

3) *Effectiveness Across DETR Variants*: Since our main experiments are based on Deformable DETR [38], we further investigate the adaptability of our method by evaluating it on different DETR variants, including DINO DETR [39] and RT DETR [35]. Notably, both DINO DETR and RT DETR incorporate a query selection mechanism that relies on ground truth labels, which contradicts the SFOD setting. To address this limitation, we disable query selection during teacher forward propagation, subsequently leveraging pseudo-labels for student query selection and SFOD training. Experimental results in Tab. VIII demonstrate that our method consistently surpasses baseline by a large margin across all three DETR variants, highlighting both its effectiveness and adaptability.

4) *Hyperparameter Sensitivity*: To evaluate the robustness of the proposed components, we conduct a hyperparameter sensitivity analysis on several key hyperparameters: (1) ω_1 and ω_2 , which balance the extra losses, (2) β and β' , which control the weights for query-based sample reweighting and feature distillation, respectively. As shown in Fig. 5, the proposed method demonstrates strong robustness to small variations in these hyperparameters, further validating its effectiveness.

5) *Contrastive Pair Construction Strategy Performance and Noise Robustness*: We evaluate two contrastive pair construction strategies in CMMB: Threshold and Matching. For Threshold, we first associate student predictions with teacher pseudo-labels by IoU-based assignment, then filter the matched pairs using a student confidence threshold. For Matching, we instead perform a global one-to-one bipartite assignment between student and teacher outputs. As shown in Fig. 4(c), the threshold strategy yields mAPs of 43.7/44.4/44.3 at thresholds 0.3/0.6/0.8, while matching achieves 44.9. This indicates that relying on student confidence introduces bias from unstable predictions, whereas bipartite matching with teacher guidance leads to more reliable contrastive pairs. We further test robustness by injecting label noise, as shown in Fig. 4(d). At 10% noise, both remain close (44.7 vs. 44.5), but as noise rises to 70%, threshold drops to 42.1 while matching

TABLE II

RESULTS OF CROSS-WEATHER ADAPTATION (FOGGY SCENARIO). “SF” DENOTES SOURCE-FREE. “V” AND “R” IN BACKBONE REPRESENT VGG AND RESNET, RESPECTIVELY. “FRCNN” STANDS FOR FASTER R-CNN. FOR EACH COLUMN, WE BOLD THE BEST AND UNDERLINE THE SECOND-BEST RESULTS SEPARATELY FOR NON-SOURCE-FREE AND SOURCE-FREE APPROACHES (EXCLUDING ORACLE)

Method	SF	Backbone	Detector	Truck	Car	Rider	Person	Train	Motor	Bicycle	Bus	mAP
Source	×	R-50	DETR	15.1	46.5	39.3	38.9	4.0	21.8	36.8	34.2	29.6
SIGMA++ [8]	×	V-16	FCOS	<u>32.1</u>	61.0	45.1	46.4	44.6	34.8	39.9	52.2	44.5
CIGAR [9]	×	V-16		27.8	62.1	47.3	46.1	44.3	33.7	41.3	56.6	44.9
EPM [10]	×	R-101		29.4	57.1	43.6	41.5	39.7	29.0	36.1	44.9	40.2
MGA-DA [11]	×	R-101		30.2	61.5	47.3	43.1	50.3	27.9	36.9	53.2	43.8
SFA [12]	×	R-50	DETR	25.1	62.6	48.6	46.5	29.4	28.3	44.0	46.2	41.3
MTTrans [13]	×	R-50		25.8	65.2	49.9	47.7	33.9	32.6	46.5	45.9	43.4
DA-DETR [14]	×	R-50		24.0	63.1	50.0	49.9	37.5	31.6	46.3	45.8	43.5
MTM [15]	×	R-50		37.2	67.2	53.4	51.0	41.6	38.4	47.7	54.4	48.9
BiADT [16]	×	R-50		31.7	<u>69.2</u>	<u>58.9</u>	<u>52.2</u>	<u>45.1</u>	<u>42.6</u>	<u>51.3</u>	<u>55.0</u>	50.8
ACCT [17]	×	R-50		31.1	69.4	58.9	53.6	<u>33.7</u>	42.6	54.4	53.5	<u>49.6</u>
SED [19]	✓	V-16	FRCNN	21.7	44.0	40.4	32.6	11.8	25.3	34.5	34.3	30.6
IRG-SFDA [20]	✓	R-50		24.4	51.9	45.2	37.4	25.2	31.5	41.6	39.6	37.1
AASFOD [22]	✓	V-16		<u>28.1</u>	44.6	44.1	32.3	29.0	31.8	38.9	34.3	35.4
BT [24]	✓	V-16		24.3	52.7	47.1	38.4	36.3	30.2	40.1	44.6	39.2
MT [27]	✓	R-50	DETR	23.5	61.5	45.0	44.1	23.9	25.0	38.0	41.2	37.8
TeST [23]	✓	R-50		27.5	54.8	46.2	45.1	11.9	32.0	44.1	44.4	38.2
DRU [26]	✓	R-50		26.2	62.5	51.5	48.3	<u>34.1</u>	34.2	48.6	43.2	<u>43.6</u>
FRANCK(ours)	✓	R-50		33.9	60.6	<u>49.3</u>	<u>48.1</u>	36.9	<u>34.0</u>	<u>47.9</u>	48.2	44.9
Oracle	✓	R-50	DETR	31.3	71.8	52.9	52.9	41.0	41.4	44.0	53.9	48.7

TABLE III

RESULTS OF CROSS-WEATHER ADAPTATION (RAINY SCENARIO). “SF” DENOTES SOURCE-FREE. “V” AND “R” IN BACKBONE REPRESENT VGG AND RESNET, RESPECTIVELY. “FRCNN” STANDS FOR FASTER R-CNN. FOR EACH COLUMN, WE BOLD THE BEST AND UNDERLINE THE SECOND-BEST RESULTS SEPARATELY FOR NON-SOURCE-FREE AND SOURCE-FREE APPROACHES (EXCLUDING ORACLE)

Method	SF	Backbone	Detector	Truck	Car	Rider	Person	Train	Motor	Bicycle	Bus	mAP
Source	×	R-50	DETR	35.0	69.0	47.4	50.9	31.5	30.5	39.9	51.0	44.4
DA-Detect [79]	×	R-50	FRCNN	38.8	61.7	47.0	40.2	47.2	34.4	29.0	<u>59.7</u>	46.0
BiADT [16]	×	R-50	DETR	41.9	75.2	52.8	54.8	42.6	36.3	47.3	61.5	51.5
AQT [80]	×	R-50		28.5	68.1	45.8	46.6	29.5	34.2	44.2	50.2	43.4
SFA [12]	×	R-50		<u>40.4</u>	<u>73.3</u>	<u>49.5</u>	<u>52.5</u>	40.1	36.6	<u>44.2</u>	52.5	<u>48.6</u>
IRG-SFDA [20]	✓	R-50	FRCNN	33.1	60.4	46.7	39.2	33.1	34.4	42.2	52.1	42.7
MT [27]	✓	R-50	DETR	34.2	73.3	49.6	51.0	39.0	29.2	45.1	48.0	46.2
DRU [26]	✓	R-50		<u>35.3</u>	73.7	51.7	52.5	<u>39.4</u>	<u>32.7</u>	<u>45.8</u>	<u>53.7</u>	<u>48.1</u>
FRANCK(ours)	✓	R-50		41.1	74.3	<u>50.9</u>	52.8	45.1	33.7	47.4	54.7	50.0
Oracle	✓	R-50	DETR	42.7	73.2	53.4	53.2	48.0	38.6	46.7	55.8	51.5

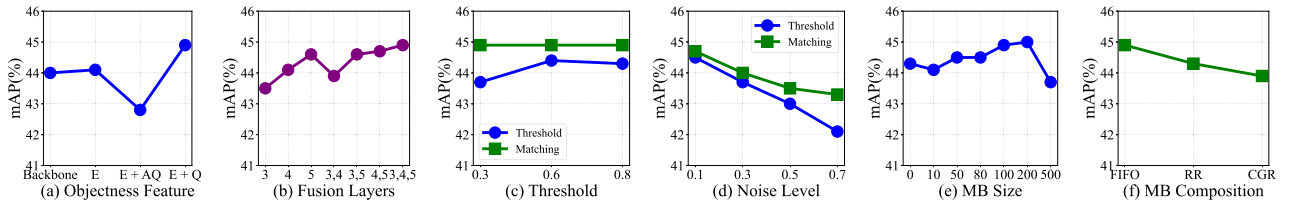


Fig. 4. Experimental results from the ablation studies: (a) Influence of different features on object estimation. (b) Influence of various encoder feature fusion layers. (c) Comparisons between thresholding and matching in CMMB. (d) Influence of controlled noise levels in CMMB. (e) Influence of memory bank (MB) size. (f) Influence of memory bank (MB) composition, including first-in-first-out (FIFO), random replacement (RR), and center-guided replacement (CGR).

stays at 43.3. This demonstrates that Hungarian matching’s global assignment mitigates mislabeled samples and maintains

a purer memory bank, whereas local thresholding admits more noisy entries.

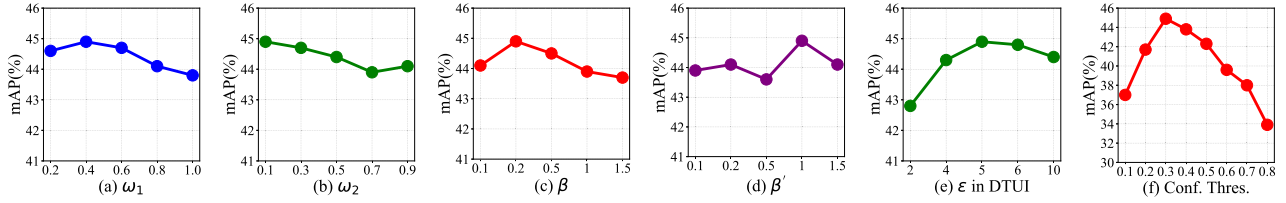


Fig. 5. Hyperparameter Sensitivity Analysis. We illustrate the sensitivity of four key hyperparameters: ω_1 , ω_2 , β , β' , ϵ , and c_{thresh} (Confidence Threshold). Each plot shows performance variation when adjusting a single hyperparameter while keeping the others fixed. The definitions of these hyperparameters are provided in Table I.

TABLE IV

RESULTS OF SYNTHETIC-TO-REAL (S2R) AND CROSS-DATASET (K2C) ADAPTATION. “SF” DENOTES SOURCE-FREE. “V” AND “R” IN BACKBONE REPRESENT VGG AND RESNET, RESPECTIVELY. “FRCNN” STANDS FOR FASTER R-CNN. FOR EACH COLUMN, WE BOLD THE BEST AND UNDERLINE THE SECOND-BEST RESULTS SEPARATELY FOR NON-SOURCE-FREE AND SOURCE-FREE APPROACHES (EXCLUDING ORACLE)

Method	SF	Backbone	Detector	mAP	
				S2R	K2C
Source	×	R-50	DETR	50.8	33.9
SFA [12]	×	R-50	DETR	52.6	46.7
DA-DETR [14]	×	R-50	DETR	54.7	48.9
IRG-SFDA [20]	✓	R-50	FRCNN	46.9	<u>45.2</u>
AASFOD [22]	✓	V-16	FRCNN	44.9	44.0
MT [27]	✓	R-50	DETR	57.0	42.5
TeST [23]	✓	R-50	DETR	57.9	42.8
DRU [26]	✓	R-50	DETR	<u>58.7</u>	45.1
FRANCK(ours)	✓	R-50	DETR	63.1	48.5
Oracle	✓	R-50	DETR	75.9	75.9

6) *Dynamic MT Updating*: Several recent studies explore dynamic teacher updates based on uncertainty estimation, leveraging techniques such as logits variance (**Var**) [26], prediction entropy (**Ent**) [63], and Soft Neighborhood Density (**SND**) [63], [97]. To further investigate this, we conduct an ablation study comparing three different updating strategies:

- 1) **Fixed intervals**. Teacher updates after constant iterations, with the meta interval defining the update period.
- 2) **Uncertainty-based updates (Var, Ent, SND)**. Updates occur when the uncertainty metric decreases, with the meta interval as the max interval, following DRU [26].
- 3) **Proposed DTUI**. The proposed updating mechanism where the meta interval corresponds to δ in Eq. (13).

Following prior work [26], we set the meta interval to 5 for uncertainty-based updates (Var, Ent, SND). As shown in Tab. X, both DTUI and uncertainty-based strategies outperform the vanilla Mean Teacher, with DTUI yielding the best results. However, uncertainty-based methods show limited gains unless combined with techniques like DRU’s student retraining [26], possibly due to noisy background queries and unstable proposals in DETR. We also observe that overly large or small intervals offer minimal benefits, highlighting the importance of a balanced update frequency in DTUI. Additionally, Fig. 5(e) shows that DTUI is robust to small

changes in ϵ , which controls interval growth. Performance only degrades when ϵ is too large (e.g., 2), delaying necessary teacher updates.

7) *Objectness Estimation Strategy*: To assess the superiority of OSSR under different objectness estimation strategies, we conduct an ablation study comparing the following approaches:

- 1) **Backbone**: Utilizing backbone features alone.
- 2) **E**: Utilizing encoder features alone.
- 3) **E+AQ**: Utilizing encoder features fused with assigned queries (i.e., queries matched with pseudo-labels via bipartite matching).
- 4) **E+Q**: Utilizing encoder features fused with all queries.

As shown in Fig. 4(a), the highest performance is achieved when using encoder features fused with assigned queries (E+AQ). This result highlights the effectiveness of objectness estimation and query-fused feature weighting. Additionally, these quantitative findings complement the visualization results in Fig. 7, further validating the impact of our approach.

8) *Pseudo-labeling Threshold*: We also conduct an ablation study on the pseudo-labeling threshold c_{thresh} , which is set to 0.3 in our main experiments. As shown in Fig. 5(f), both excessively low and high thresholds lead to suboptimal detection performance, while a threshold around 0.3 yields the best results. This finding is consistent with DRU [26], which also uses a 0.3 threshold for DETR-based pseudo-labeling.

9) *Memory Bank in CMMB*: We conduct ablations on both memory bank size and update strategy. Sizes range from 0 (disabling the memory bank) to several capacities per class, and update strategies include FIFO, random replacement (RR), and center-guided replacement (CGR), where a new feature replaces the entry farthest from the current class center. As shown in Fig. 4(e) and (f), moderate sizes (around 100–200 entries) give the best mAP, while very small or very large sizes slightly reduce performance. Among update strategies, FIFO performs best, with RR and CGR yielding slightly lower results. These findings suggest that a balanced memory size and stable FIFO updates provide the most effective composition for contrastive learning in CMMB.

10) *Multi-scale Encoder Feature Fusion*: To assess the impact of multi-scale encoder feature fusion in OSSR, we conduct an ablation study on different fusion strategies using layers 3, 4, 5. We experiment with fusing features from individual layers as well as multiple layers. As shown in Fig. 4(c), (1) reweighting based on lower-layer encoder features leads to suboptimal performance, as they emphasize high-level

TABLE V

RESULTS OF CROSS-SCENE ADAPTATION. “SF” DENOTES SOURCE-FREE. “V” AND “R” IN BACKBONE REPRESENT VGG AND RESNET, RESPECTIVELY. “FRCNN” STANDS FOR FASTER R-CNN. FOR EACH COLUMN, WE BOLD THE BEST AND UNDERLINE THE SECOND-BEST RESULTS SEPARATELY FOR NON-SOURCE-FREE AND SOURCE-FREE APPROACHES (EXCLUDING ORACLE)

Method	SF	Backbone	Detector	Truck	Car	Rider	Person	Motor	Bicycle	Bus	mAP
Source	×	R-50	DETR	17.5	57.0	29.4	43.7	15.6	17.7	17.6	28.3
SIGMA++ [8]	×	V-16	FCOS	21.1	<u>65.6</u>	30.4	47.5	27.1	17.8	26.3	33.7
EPM [10]	×	V-16		18.8	55.8	26.8	39.6	20.1	14.5	19.1	27.8
SFA [12]	×	R-50	DETR	19.1	57.5	27.6	40.2	19.2	15.4	23.4	28.9
MTTrans [13]	×	R-50		<u>25.1</u>	61.5	30.1	44.1	23.0	17.7	<u>26.9</u>	32.6
MTM [15]	×	R-50		23.0	68.8	<u>35.1</u>	53.7	<u>28.0</u>	<u>23.8</u>	28.8	<u>37.3</u>
BiADT [16]	×	R-50		17.4	60.9	34.0	42.1	25.7	18.2	19.5	31.1
ACCT [17]	×	R-50		26.0	61.8	41.4	<u>51.8</u>	36.9	31.7	23.4	39.0
SED [19]	✓	V-16	FRCNN	20.6	50.4	32.6	32.4	25.0	18.9	23.4	29.0
AASFOD [22]	✓	V-16		26.6	50.2	36.3	33.2	22.5	28.2	24.4	31.6
BT [24]	✓	V-16		24.2	50.4	34.6	32.7	28.5	24.7	24.9	31.4
MT [27]	✓	R-50	DETR	21.8	63.0	36.0	50.5	20.9	22.0	22.4	33.8
TeST [23]	✓	R-50		23.2	<u>63.8</u>	<u>37.5</u>	<u>51.7</u>	22.7	23.4	23.6	35.1
DRU [26]	✓	R-50		<u>27.1</u>	<u>62.7</u>	36.9	45.8	32.5	22.7	<u>28.1</u>	<u>36.6</u>
FRANCK(ours)	✓	R-50		29.5	65.6	43.3	55.0	<u>30.0</u>	<u>28.0</u>	33.6	40.7
Oracle	✓	R-50	DETR	66.9	87.9	56.4	74.9	53.8	68.3	55.0	66.2

TABLE VI

RESULTS OF CITYSCAPES-TO-ACDC ADAPTATION ON FOUR DIVERSE WEATHER CONDITIONS. “SF” DENOTES SOURCE-FREE. “FRCNN” STANDS FOR FASTER R-CNN. WE BOLD THE BEST AND UNDERLINE THE SECOND-BEST RESULTS FOR SOURCE-FREE APPROACHES

Method	SF	Detector	Snow	Rain	Night	Fog
AT [91]	×	FRCNN	55.2	37.7	29.5	62.2
DT-G [92]	×		56.8	39.0	36.4	68.6
DRU [26]	✓	DETR	<u>37.9</u>	<u>26.3</u>	<u>16.5</u>	45.4
FRANCK(ours)	✓		42.4	29.9	17.5	49.0

semantics over object-level attention. While fusing layer 5 with others yields only marginal gains, it confirms that multi-layer fusion provides more comprehensive features than a single layer, ultimately enhancing overall performance.

1) *Effectiveness Over Different Detectors*: To further illustrate the DETR-specific design, we directly transfer our method to Faster R-CNN [3] by replacing query and encoder features with RoI and backbone features. As shown in Tab. XI, the modules yield only marginal or even degraded gains compared to the clear improvements on DETR. This stems from our DETR-specific design: OSSR and UQFD rely on query–encoder fused objectness, while RoI features lack the global semantic context of DETR queries, weakening supervision and distillation; CMMB also underperforms as it is built around learnable queries rather than local RoIs. Moreover, Hungarian matching on thousands of RPN proposals incurs extra cost and instability. These findings underscore the query-centric nature of our approach and suggest that a promising direction for future work is to generalize the key ideas and designs of DETR-based SFOD so that they can be effectively adapted to a broader range of detector architectures.

F. Visualization Study

To further illustrate the effectiveness of our method, we present a visualization study in this section.

1) *Detection Result Visualization*: To showcase effective domain adaptation, we visualize detection results from three settings: (1) **Source-only**, (2) **FRANCK**, and (3) **Ground Truth**. As shown in Fig. 6, our method significantly improves both object localization and classification in the target domain, reducing false positives and enhancing detection quality.

2) *Object Attention Visualization*: To illustrate the impact of query-fused objectness scores on encoder features, we compare OSSR objectness attention generation methods, including “E”, “E+AQ”, and “E+Q”, as outlined previously. To better visualize the weight distribution, we also subtract the scaled attention score of “E+Q” from 1, since Eq. (8) assigns **high** weights to **low** attention scores.

As shown in Fig. 7, domain shifts often lead encoder features to over-focus on background regions. Query fusion mitigates this by redirecting attention toward objects, with “E+AQ” showing the most improvement. However, treating all queries equally (“E+Q”) can suppress attention to foreground and hard-to-detect objects, consistent with findings in DETRDistill [72]. Our reweighting strategy corrects this by assigning higher weights to under-recognized areas, addressing both foreground-background and easy-hard sample imbalances. This improves feature discriminability, as further validated in Fig. 4(a).

G. Limitation

While our method achieves state-of-the-art performance, it has limitations that point to future directions. First, although a lower confidence threshold works well across varied scenarios, it may underperform in precision-critical cases. A dynamic

TABLE VII

PERFORMANCE COMPARISON OF DIFFERENT BACKBONES. WE USE DEFORMABLE DETR AS THE BASE DETECTOR. WE BOLD THE BEST AND UNDERLINE THE SECOND-BEST RESULTS SEPARATELY FOR EACH BACKBONE

Method	Backbone	Truck	Car	Rider	Person	Train	Motor	Bicycle	Bus	mAP
Source	R-50 [2]	15.1	46.5	39.3	38.9	4.0	21.8	36.8	34.2	29.6
MT [27]		23.5	<u>61.5</u>	45.0	44.1	23.9	25.0	38.0	41.2	37.8
DRU [26]		<u>26.2</u>	62.5	51.5	48.3	34.1	34.2	48.6	43.2	43.6
FRANCK(ours)		33.9	60.6	<u>49.3</u>	<u>48.1</u>	36.9	<u>34.0</u>	<u>47.9</u>	48.2	44.9
Source	ViT-B [94]	17.8	46.8	41.3	38.2	3.7	24.2	37.2	30.1	29.9
MT [27]		22.3	62.0	46.1	<u>45.9</u>	11.2	<u>27.5</u>	41.1	37.7	36.7
DRU [26]		<u>22.5</u>	62.6	45.9	44.8	20.9	25.5	<u>41.1</u>	<u>39.2</u>	<u>37.8</u>
FRANCK(ours)		25.7	<u>62.2</u>	49.6	46.7	<u>14.1</u>	30.4	42.0	43.3	39.3
Source	Swin-T [95]	30.0	55.5	48.5	45.0	15.8	35.3	42.6	41.3	39.2
MT [27]		33.3	<u>65.5</u>	48.9	47.2	19.7	34.4	46.9	40.9	42.1
DRU [26]		<u>35.0</u>	66.4	<u>49.7</u>	<u>49.8</u>	30.4	<u>38.1</u>	47.6	46.9	45.5
FRANCK(ours)		35.2	65.0	52.3	50.5	<u>28.4</u>	41.0	50.0	47.2	46.2
Source	Swin-S [95]	29.9	56.2	50.5	45.6	35.9	36.6	42.7	46.6	43.0
MT [27]		36.6	66.9	52.1	49.1	25.3	40.2	47.2	47.0	45.6
DRU [26]		37.8	65.3	<u>53.8</u>	<u>50.8</u>	<u>27.6</u>	<u>41.8</u>	<u>49.2</u>	<u>50.7</u>	<u>47.1</u>
FRANCK(ours)		<u>37.3</u>	<u>65.8</u>	54.0	51.4	26.9	42.3	50.1	52.1	47.5
Source	Swin-B [95]	37.7	62.5	51.3	49.3	39.2	38.6	45.2	46.2	46.2
MT [27]		<u>41.2</u>	67.2	48.6	46.4	39.0	<u>43.2</u>	46.8	49.7	47.9
DRU [26]		39.4	<u>69.1</u>	<u>52.5</u>	51.8	42.1	44.7	50.8	<u>56.4</u>	<u>50.9</u>
FRANCK(ours)		43.1	69.2	53.4	53.1	42.1	42.7	52.6	56.8	51.6
Source	Swin-L [95]	39.4	63.5	49.5	49.2	48.9	41.9	45.6	52.7	48.9
MT [27]		44.0	68.0	51.5	49.7	42.2	43.6	47.3	55.2	49.5
DRU [26]		<u>45.0</u>	<u>68.8</u>	<u>52.9</u>	<u>53.1</u>	47.8	<u>45.2</u>	<u>50.6</u>	<u>56.6</u>	<u>52.5</u>
FRANCK(ours)		46.0	69.4	55.8	54.4	<u>46.1</u>	48.2	51.2	59.1	53.8

TABLE VIII

PERFORMANCE COMPARISON OF DIFFERENT DETR VARIANTS. WE USE RESNET-50 AS THE BACKBONE. WE BOLD THE BEST AND UNDERLINE THE SECOND-BEST RESULTS SEPARATELY FOR EACH KIND OF DETECTOR

Method	Detector	Truck	Car	Rider	Person	Train	Motor	Bicycle	Bus	mAP
Source	Deformable DETR [38]	15.1	46.5	39.3	38.9	4.0	21.8	36.8	34.2	29.6
MT [27]		<u>23.5</u>	61.5	<u>45.0</u>	<u>44.1</u>	<u>23.9</u>	<u>25.0</u>	<u>38.0</u>	<u>41.2</u>	<u>37.8</u>
FRANCK(ours)		33.9	<u>60.6</u>	49.3	48.1	36.9	34.0	47.9	48.2	44.9
Source	DINO DETR [39]	<u>18.9</u>	46.1	42.0	39.5	7.5	24.1	40.1	35.1	31.7
MT [27]		18.5	<u>64.7</u>	<u>46.0</u>	<u>49.5</u>	<u>37.8</u>	<u>35.6</u>	<u>44.8</u>	<u>44.7</u>	<u>42.7</u>
FRANCK(ours)		29.2	65.0	54.3	51.8	40.3	35.6	50.6	49.7	47.1
Source	RT DETR [35]	29.9	54.0	46.8	38.7	21.9	29.4	39.3	43.7	38.0
MT [27]		<u>34.5</u>	<u>64.5</u>	<u>51.2</u>	<u>45.0</u>	<u>37.4</u>	<u>36.3</u>	<u>43.4</u>	<u>54.6</u>	<u>45.9</u>
FRANCK(ours)		44.2	73.4	54.8	51.3	52.1	42.9	48.6	65.8	54.1

TABLE IX

ABLATION STUDY ON COMPONENT ANALYSIS. THE REPORTED *Time* DENOTES THE PER-ITERATION PROCESSING TIME, AND $\Delta Time$ INDICATES THE RELATIVE INCREASE COMPARED WITH THE MT+DTUI BASELINE

MT+DTUI	CMMB	OSSR	UQFD	mAP	Time/ms	$\Delta Time/\%$
✓				41.7	387.5	–
✓			✓	42.1	508.7	23.8
✓		✓		42.3	428.8	9.6
✓	✓			43.5	416.2	6.9
✓		✓	✓	43.2	517.7	25.1
✓	✓		✓	44.2	519.9	25.5
✓	✓	✓		44.4	477.2	18.8
✓	✓	✓	✓	44.9	542.1	28.5

TABLE X

ABLATION STUDY ON DYNAMIC MT(MEAN TEACHER) UPDATING ANALYSIS. “Var”, “Ent”, and “SND” DENOTE LOGITS VARIANCE, PREDICTION ENTROPY, AND SOFT NEIGHBORHOOD DENSITY, RESPECTIVELY, AS STATED IN SEC. IV-E. MI REPRESENTS META INTERVAL

Strategy	Fixed intervals				Var	Ent	SND	DTUI		
MI	1	2	5	10	5	5	5	2	5	10
mAP	41.4	42.4	44.1	41.7	44.4	43.6	43.8	43.5	44.9	43.1

threshold, starting low to mine potential objects and increasing later to enhance precision, could better balance recall and

precision. Second, with the rise of vision foundation models (VFMs) like CLIP [98], incorporating VFM-guided cues (e.g., text-image similarity for refining memory bank samples) could



Fig. 6. Visualization of detection results before and after adaptation and gt labels. From top to bottom: Source-only results, FRANCK results, and ground truth. We show both Cross-weather adaptation results and Synthetic-to-real adaptation results in this experiment.

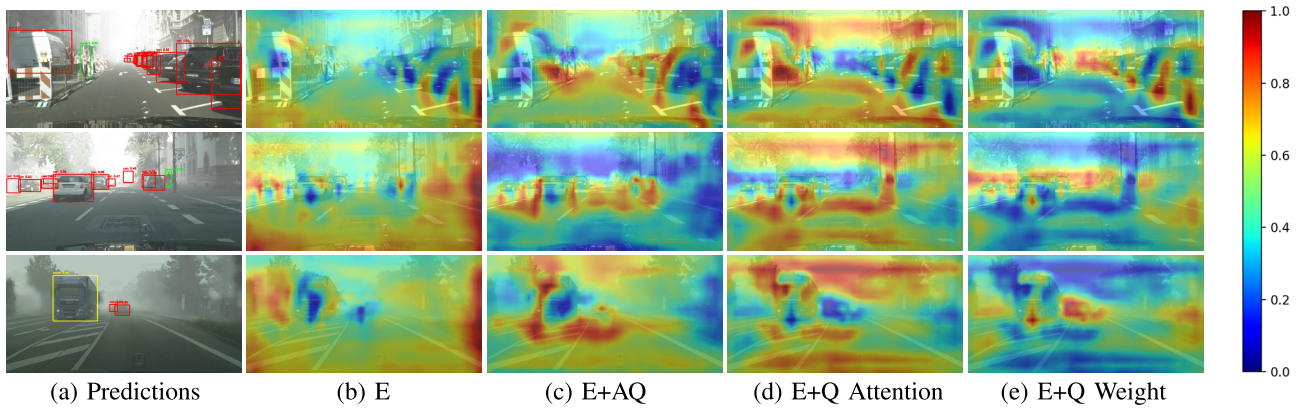


Fig. 7. Visualization of object attention visualization with different strategies. “E” is for Encoder feature only, “E+AQ” is for fusion of Encoder feature and assigned query features, and “E+Q” is for fusion of Encoder feature and all query features. In the proposed method, we assign **high** weights to **low** E+Q attention scores and vice versa.

TABLE XI

ABLATION STUDY ON DIFFERENT MODULE COMBINATIONS OVER DIFFERENT DETECTORS

Detector	MT+DTUI	CMMB	OSSR	UQFD	mAP
FRCNN	✓				37.4
	✓	✓			37.9
	✓		✓		36.4
	✓			✓	37.1
	✓	✓	✓	✓	37.2
DETR	✓	✓	✓	✓	44.9

further enhance contrastive learning, especially for applications less constrained by computation.

V. CONCLUSION

In this paper, we tackle the challenge of source-free domain adaptive object detection (SFOD), specifically focusing on adapting source-pre-trained DETR networks to target domains without access to source data. To this end, we propose FRANCK, a novel framework that fully exploits DETR-specific features by incorporating four key components: (1) an Objectness Score-based Sample Reweighting (OSSR) module, (2) a Contrastive Learning with Matching-based Memory Bank (CMMB) module, (3) an Uncertainty-weighted Query-fused

Feature Distillation (UQFD) module, and (4) an enhanced self-training pipeline with Dynamic Teacher Updating Interval (DTUI). Our method achieves state-of-the-art performance, surpassing previous SFOD approaches across multiple widely used benchmarks. For future work, we aim to extend our framework to more realistic scenarios, such as multi-source adaptation. We hope this work provides valuable insights and inspiration for advancing DAOD and SFOD, further contributing to the broader research community.

REFERENCES

- [1] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “Imagenet classification with deep convolutional neural networks,” in *Proc. Adv. Neural Inf. Process. Syst.*, 2012, pp. 1106–1114.
- [2] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.
- [3] S. Ren, K. He, R. Girshick, and J. Sun, “Faster R-CNN: Towards real-time object detection with region proposal networks,” in *Proc. Int. Conf. Adv. Neural Inf. Process. Syst.*, vol. 28, 2015, pp. 91–99.
- [4] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, “End-to-end object detection with transformers,” in *Proc. Eur. Conf. Comput. Vis. Cham, Switzerland: Springer*, 2020, pp. 213–229.
- [5] D. Amodei, C. Olah, J. Steinhardt, P. Christiano, J. Schulman, and D. Mané, “Concrete problems in AI safety,” 2016, *arXiv:1606.06565*.
- [6] A. L. Rodriguez and K. Mikołajczyk, “Domain adaptation for object detection via style consistency,” 2019, *arXiv:1911.10033*.

- [7] G. Mattolin, L. Zanella, E. Ricci, and Y. Wang, "ConfMix: Unsupervised domain adaptation for object detection via confidence-based mixing," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2023, pp. 423–433.
- [8] W. Li, X. Liu, and Y. Yuan, "SIGMA++: Improved semantic-complete graph matching for domain adaptive object detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 7, pp. 9022–9040, Jul. 2023.
- [9] Y. Liu, J. Wang, C. Huang, Y. Wang, and Y. Xu, "CIGAR: Cross-modality graph reasoning for domain adaptive object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 23776–23786.
- [10] C.-C. Hsu, Y.-H. Tsai, Y.-Y. Lin, and M.-H. Yang, "Every pixel matters: Center-aware feature alignment for domain adaptive object detector," in *Proc. Eur. Conf. Comput. Vis. Cham, Switzerland: Springer*, 2020, pp. 733–748.
- [11] W. Zhou, D. Du, L. Zhang, T. Luo, and Y. Wu, "Multi-granularity alignment domain adaptation for object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 9581–9590.
- [12] W. Wang et al., "Exploring sequence feature alignment for domain adaptive detection transformers," in *Proc. 29th ACM Int. Conf. Multimedia*, Oct. 2021, pp. 1730–1738.
- [13] J. Yu et al., "MTTrans: Cross-domain object detection with mean teacher transformer," in *Proc. 17th Eur. Conf. Comput. Vis. Cham, Switzerland: Springer*, 2022, pp. 629–645.
- [14] J. Zhang, J. Huang, Z. Luo, G. Zhang, X. Zhang, and S. Lu, "DA-DETR: Domain adaptive detection transformer with information fusion," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 23787–23798.
- [15] W. Weng and C. Yuan, "Mean teacher DETR with masked feature alignment: A robust domain adaptive detection transformer framework," in *Proc. AAAI Conf. Artif. Intell.*, 2024, vol. 38, no. 6, pp. 5912–5920.
- [16] L. He, W. Wang, A. Chen, M. Sun, C.-H. Kuo, and S. Todorovic, "Bidirectional alignment for domain adaptive detection with transformers," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2023, pp. 18729–18739.
- [17] Z. Zeng, Y. Ding, and H. Lu, "Enhancing cross-domain detection: Adaptive class-aware contrastive transformer," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Apr. 2024, pp. 6670–6674.
- [18] J. Liang, R. He, and T. Tan, "A comprehensive survey on test-time adaptation under distribution shifts," *Int. J. Comput. Vis.*, vol. 133, no. 1, pp. 1–34, Jan. 2025.
- [19] X. Li et al., "A free lunch for unsupervised domain adaptive object detection without source data," in *Proc. AAAI Conf. Artif. Intell.*, Aug. 2021, pp. 8474–8481.
- [20] V. Vs, P. Oza, and V. M. Patel, "Instance relation graph guided source-free domain adaptive object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 3520–3530.
- [21] S. Li, M. Ye, X. Zhu, L. Zhou, and L. Xiong, "Source-free object detection by learning to overlook domain style," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 8014–8023.
- [22] Q. Chu, S. Li, G. Chen, K. Li, and X. Li, "Adversarial alignment for source free object detection," in *Proc. AAAI Conf. Artif. Intell.*, Jun. 2023, pp. 452–460.
- [23] S. Sinha, P. Gehler, F. Locatello, and B. Schiele, "TeST: Test-time self-training under distribution shift," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2023, pp. 2758–2768.
- [24] J. Deng, W. Li, and L. Duan, "Balanced teacher for source-free object detection," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 34, no. 8, pp. 7231–7243, Aug. 2024.
- [25] S. Zhao, H. Yao, C. Lin, Y. Gao, and G. Ding, "Multi-source-free domain adaptive object detection," *Int. J. Comput. Vis.*, vol. 132, no. 12, pp. 5950–5982, Dec. 2024.
- [26] T. Le Ba Khanh, H.-H. Nguyen, L. H. Pham, D. N.-N. Tran, and J. W. Jeon, "Dynamic retraining-updating mean teacher for source-free object detection," 2024, *arXiv:2407.16497*.
- [27] A. Tarvainen and H. Valpola, "Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results," in *Proc. Int. Conf. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 1195–1204.
- [28] Z. Chen, Z. Wang, and Y. Zhang, "Exploiting low-confidence pseudo-labels for source-free object detection," in *Proc. 31st ACM Int. Conf. Multimedia*, Oct. 2023, pp. 5370–5379.
- [29] S. Suzuki, T. Yamane, N. Makishima, K. Suzuki, A. Ando, and R. Masumura, "OnDA-DETR: Online domain adaptation for detection transformers with self-training framework," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Oct. 2023, pp. 1780–1785.
- [30] R. Girshick, "Fast R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 1440–1448.
- [31] W. Liu et al., "SSD: Single shot multibox detector," in *Proc. 14th Eur. Conf. Amsterdam, The Netherlands: Springer*, Oct. 2016, pp. 21–37.
- [32] Z. Tian, C. Shen, H. Chen, and T. He, "FCOS: Fully convolutional one-stage object detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 9627–9636.
- [33] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 779–788.
- [34] D. Jia et al., "DETRs with hybrid matching," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 19702–19712.
- [35] Y. Zhao et al., "DETRs beat YOLOs on real-time object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2024, pp. 16965–16974.
- [36] S. Liu et al., "DAB-DETR: Dynamic anchor boxes are better queries for DETR," 2022, *arXiv:2201.12329*.
- [37] F. Li, H. Zhang, S. Liu, J. Guo, L. M. Ni, and L. Zhang, "DN-DETR: Accelerate DETR training by introducing Query DeNoising," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 13619–13627.
- [38] X. Zhu, W. Su, L. Lu, B. Li, X. Wang, and J. Dai, "Deformable DETR: Deformable transformers for end-to-end object detection," in *Proc. Int. Conf. Learn. Represent.*, 2020.
- [39] H. Zhang et al., "DINO: DETR with improved DeNoising anchor boxes for end-to-end object detection," 2022, *arXiv:2203.03605*.
- [40] Y. Chen, W. Li, C. Sakaridis, D. Dai, and L. Van Gool, "Domain adaptive faster R-CNN for object detection in the wild," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2018, pp. 3339–3348.
- [41] J. Deng, D. Xu, W. Li, and L. Duan, "Harmonious teacher for cross-domain object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 23829–23838.
- [42] Y. Zhang, Z. Wang, and Y. Mao, "RPN prototype alignment for domain adaptive object detector," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 12420–12429.
- [43] Z. He and L. Zhang, "Domain adaptive object detection via asymmetric tri-way faster-RCNN," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 309–324.
- [44] X. Yao, S. Zhao, P. Xu, and J. Yang, "Multi-source domain adaptation for object detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 3273–3282.
- [45] S. Cao, D. Joshi, L.-Y. Gui, and Y.-X. Wang, "Contrastive mean teacher for domain adaptive object detectors," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 23839–23848.
- [46] V. S. Vibashan et al., "MeGA-CDA: Memory guided attention for category-aware unsupervised domain adaptive object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2021, pp. 4516–4526.
- [47] P. Li, Y. He, F. R. Yu, P. Song, D. Yin, and G. Zhou, "IGG: Improved graph generation for domain adaptive object detection," in *Proc. 31st ACM Int. Conf. Multimedia*, Oct. 2023, pp. 1314–1324.
- [48] H. Li et al., "REACT: Remainder adaptive compensation for domain adaptive object detection," *IEEE Trans. Image Process.*, vol. 33, pp. 3735–3748, 2024.
- [49] S. Zhang, H. Tuo, J. Hu, and Z. Jing, "Domain adaptive YOLO for one-stage cross-domain detection," in *Proc. Asian Conf. Mach. Learn.*, Apr. 2021, pp. 785–797.
- [50] M. Hnewa and H. Radha, "Integrated multiscale domain adaptive YOLO," *IEEE Trans. Image Process.*, vol. 32, pp. 1857–1867, 2023.
- [51] M. Jing, J. Li, K. Lu, L. Zhu, and H. T. Shen, "Visually source-free domain adaptation via adversarial style matching," *IEEE Trans. Image Process.*, vol. 33, pp. 1032–1044, 2024.
- [52] Y. Fang, P.-T. Yap, W. Lin, H. Zhu, and M. Liu, "Source-free unsupervised domain adaptation: A survey," *Neural Netw.*, vol. 174, Jun. 2024, Art. no. 106230.
- [53] J. Liang, D. Hu, and J. Feng, "Do we really need to access the source data? Source hypothesis transfer for unsupervised domain adaptation," in *Proc. Int. Conf. Mach. Learn.*, Feb. 2020, pp. 6028–6039.
- [54] Y. Zhang, Z. Wang, and W. He, "Class relationship embedded learning for source-free unsupervised domain adaptation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 7619–7629.
- [55] Y. Liu, W. Zhang, and J. Wang, "Source-free domain adaptation for semantic segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 1215–1224.

- [56] Q. Peng, C. Zheng, and C. Chen, "Source-free domain adaptive human pose estimation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2023, pp. 4803–4813.
- [57] W. Guo, Y. Sun, Y. Xu, Z. Qiao, Y. Yang, and H. Xiong, "SpGesture: Source-free domain-adaptive sEMG-based gesture recognition with Jacard attentive spiking neural network," 2024, *arXiv:2405.14398*.
- [58] X. Zheng, P. Y. Zhou, A. V. Vasilakos, and L. Wang, "360SFUDA++: Towards source-free UDA for panoramic segmentation by learning reliable category prototypes," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 47, no. 2, pp. 1190–1204, Feb. 2025.
- [59] H. Xia, H. Zhao, and Z. Ding, "Adaptive adversarial network for source-free domain adaptation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 9010–9019.
- [60] P. Agarwal, D. P. Paudel, J.-N. Zaech, and L. Van Gool, "Unsupervised robust domain adaptation without source data," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2022, pp. 2805–2814.
- [61] A. Bochkovskiy, C.-Y. Wang, and H.-Y. M. Liao, "YOLOv4: Optimal speed and accuracy of object detection," 2020, *arXiv:2004.10934*.
- [62] L. Hoyer, D. Dai, H. Wang, and L. Van Gool, "MIC: Masked image consistency for context-enhanced domain adaptation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 11721–11732.
- [63] D. Zhao, S. Wang, Q. Zang, D. Quan, X. Ye, and L. Jiao, "Towards better stability and adaptability: Improve online self-training for model adaptation in semantic segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 11733–11743.
- [64] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2980–2988.
- [65] P. Khosla et al., "Supervised contrastive learning," in *Proc. NIPS*, Mar. 2020, pp. 18661–18673.
- [66] Z. Wu, Y. Xiong, S. X. Yu, and D. Lin, "Unsupervised feature learning via non-parametric instance discrimination," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 3733–3742.
- [67] Y. Tian, D. Krishnan, and P. Isola, "Contrastive multiview coding," in *Proc. Eur. Conf. Comput. Vis. Cham, Switzerland: Springer*, 2020, pp. 776–794.
- [68] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2961–2969.
- [69] X. Li et al., "Generalized focal loss: Learning qualified and distributed bounding boxes for dense object detection," in *Proc. Annu. Conf. Neural Inf. Process. Syst. (NIPS)*, H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, Eds., Dec. 2020, pp. 21002–21012cc/paper/2020/hash/f0bda020d2470f2e74990a07a607ebd9-Abstract.html.
- [70] Z. Wang, X. Xie, J. Yang, and G. Shi, "Soft focal loss: Evaluating sample quality for dense object detection," *Neurocomputing*, vol. 480, pp. 271–280, Apr. 2022.
- [71] A. Gupta, S. Narayan, K. J. Joseph, S. Khan, F. S. Khan, and M. Shah, "OW-DETR: Open-world detection transformer," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 9235–9244.
- [72] J. Chang, S. Wang, H.-M. Xu, Z. Chen, C. Yang, and F. Zhao, "DETRDistill: A universal knowledge distillation framework for DETR-families," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2023, pp. 6875–6885.
- [73] S. S. Mullapally, A. S. Gehlot, R. M. Anwer, F. S. Khan, and H. Cholakkal, "Semi-supervised open-world object detection," in *Proc. AAAI Conf. Artif. Intell.*, 2024, vol. 38, no. 5, pp. 4305–4314.
- [74] M. Jaderberg, K. Simonyan, A. Zisserman, and K. Kavukcuoglu, "Spatial transformer networks," in *Proc. Adv. Neural Inf. Process. Syst. Annu. Conf. Neural Inf. Process. Syst.*, vol. 28, Dec. 2015, pp. 2017–2025.
- [75] K. Li et al., "Improving cross-domain detection with self-supervised learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2023, pp. 4746–4755.
- [76] T. Wang, L. Yuan, X. Zhang, and J. Feng, "Distilling object detectors with fine-grained feature imitation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2019, pp. 4933–4942.
- [77] B. Zhao, Q. Cui, R. Song, Y. Qiu, and J. Liang, "Decoupled knowledge distillation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 11953–11962.
- [78] Z. Chen, C. Yang, Q. Li, F. Zhao, Z.-J. Zha, and F. Wu, "Disentangle your dense object detector," in *Proc. 29th ACM Int. Conf. Multimedia*, Oct. 2021, pp. 4939–4948.
- [79] J. Li et al., "Domain adaptation based object detection for autonomous driving in foggy and rainy weather," 2023, *arXiv:2307.09676*.
- [80] W.-J. Huang, Y.-L. Lu, S.-Y. Lin, Y. Xie, and Y.-Y. Lin, "AQT: Adversarial Query transformers for domain adaptive object detection," in *Proc. 31st Int. Joint Conf. Artif. Intell.*, Jul. 2022, pp. 972–979.
- [81] M. Cordts et al., "The cityscapes dataset for semantic urban scene understanding," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 3213–3223.
- [82] C. Sakaridis, D. Dai, and L. Van Gool, "Semantic foggy scene understanding with synthetic data," *Int. J. Comput. Vis.*, vol. 126, no. 9, pp. 973–992, Sep. 2018.
- [83] M. Johnson-Roberson, C. Barto, R. Mehta, S. N. Sridhar, K. Rosaen, and R. Vasudevan, "Driving in the matrix: Can virtual worlds replace human-generated annotations for real world tasks?," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, May 2017, pp. 746–753.
- [84] F. Yu et al., "BDD100K: A diverse driving dataset for heterogeneous multitask learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 2633–2642.
- [85] Q. Guo et al., "EfficientDeRain: Learning pixel-wise dilation filtering for high-efficiency single-image deraining," in *Proc. AAAI Conf. Artif. Intell.*, May 2021, pp. 1487–1495.
- [86] J. Li, R. Xu, J. Ma, Q. Zou, J. Ma, and H. Yu, "Domain adaptive object detection for autonomous driving under foggy weather," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2023, pp. 612–622.
- [87] X. Hu, C.-W. Fu, L. Zhu, and P.-A. Heng, "Depth-attentional features for single-image rain removal," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 8022–8031.
- [88] V. Sindagi, P. Oza, R. Yasarla, and V. Patel, "Prior-based domain adaptive object detection for hazy and rainy conditions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 763–780.
- [89] A. Geiger, P. Lenz, and R. Urtasun, "Are we ready for autonomous driving? The KITTI vision benchmark suite," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2012, pp. 3354–3361.
- [90] C. Sakaridis, D. Dai, and L. Van Gool, "ACDC: The adverse conditions dataset with correspondences for semantic driving scene understanding," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 10765–10775.
- [91] Y.-J. Li et al., "Cross-domain adaptive teacher for object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 7581–7590.
- [92] M.-A. Lavoie, A. Mahmoud, and S. L. Waslander, "Large self-supervised models bridge the gap in domain adaptive object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2025, pp. 4692–4702.
- [93] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*.
- [94] A. Dosovitskiy et al., "An image is worth 16×16 words: Transformers for image recognition at scale," 2020, *arXiv:2010.11929*.
- [95] Z. Liu et al., "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 10012–10022.
- [96] W. Wang et al., "Pyramid vision transformer: A versatile backbone for dense prediction without convolutions," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 568–578.
- [97] K. Saito, D. Kim, P. Teterwak, S. Sclaroff, T. Darrell, and K. Saenko, "Tune it the right way: Unsupervised validation of domain adaptation via soft neighborhood density," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 9164–9173.
- [98] A. Radford et al., "Learning transferable visual models from natural language supervision," in *Proc. Int. Conf. Mach. Learn.*, vol. 139, 2021, pp. 8748–8763.



Huizai Yao received the B.Eng. degree from Nankai University, Tianjin, China, in 2024. He is currently a Research Intern with BNRist, Tsinghua University, and Shanghai Artificial Intelligence Laboratory. His research interests include deep learning, transfer learning, and computer vision.



Sicheng Zhao (Senior Member, IEEE) received the Ph.D. degree from Harbin Institute of Technology, Harbin, China, in 2016. He was a Visiting Scholar with the National University of Singapore, Singapore, from 2013 to 2014; a Research Fellow with Tsinghua University, Beijing, China, from 2016 to 2017; a Postdoctoral Research Fellow with the University of California at Berkeley, Berkeley, CA, USA, from 2017 to 2020; and a Postdoctoral Research Scientist with Columbia University, New York, NY, USA, from 2020 to 2022. He is currently a

Research Associate Professor with Tsinghua University. His research interests include affective computing, multimedia, and computer vision. He is an Associate Editor of IEEE TRANSACTIONS ON IMAGE PROCESSING and IEEE TRANSACTIONS ON AFFECTIVE COMPUTING.



Tengfei Xing received the M.S. degree from the University of Science and Technology Beijing in 2012. He is currently a Senior Expert Algorithm Engineer with Didi Chuxing. His research interests include computer vision and multi-modal fusion in intelligent transportation.



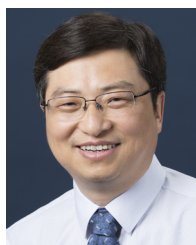
Shuo Lu received the B.S. degree from Harbin Institute of Technology, Harbin, China, in 2024. He is currently pursuing the master's degree with the Institute of Automation, Chinese Academy of Sciences. His research interests include affective computing and trustworthy artificial intelligence.



Chenggang Yan received the B.S. degree in control science and engineering from Shandong University, Shandong, China, in 2008, and the Ph.D. degree in computer science from the University of Chinese Academy of Sciences, Beijing, China, in 2013. He is currently a Professor with the Department of Automation, Hangzhou Dianzi University. His research interests include computational photography, pattern recognition, and intelligent systems.



Hui Chen is currently an Assistant Researcher with Beijing National Research Center for Information Science and Technology, Tsinghua University. He has published more than 15 top peer-reviewed conferences and journals, including CVPR, ICCV, and ICLR. His research interests include efficient and effective multi-modal perception and learning. He served as a PC member for several top-tier conferences.



Jianhua Tao (Senior Member, IEEE) received the Ph.D. degree from Tsinghua University, China, in 2001. He has directed many national projects, including the "863" National Natural Science Foundation of China. He has published more than 80 papers in journals and proceedings, including IEEE TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING, ICASSP, and INTER-SPEECH. His research interests include speech synthesis, affective computing, and pattern recognition. He serves as the Steering Committee Member

for IEEE TRANSACTIONS ON AFFECTIVE COMPUTING and the Chair or a Program Committee Member for major conferences, including ICPR and Interspeech. He is the Winner of the National Science Fund for Distinguished Young Scholars and the Deputy Director of NLPR, CASIA.



Yangyang Li received the Ph.D. degree in computer science from Beijing University of Posts and Telecommunications in 2015. He was a Senior Engineer with the National Engineering Research Center for Risk Perception and Prevention from 2015 to 2022. He is currently a Professor-Level Senior Engineer with the Academy of Cyber. His research interests include cyberspace security, social networks, and data science.



Guiguang Ding (Senior Member, IEEE) received the Ph.D. degree from Xidian University, China, in 2004. He is currently a Professor with the School of Software, Tsinghua University. Before joining the School of Software in 2006, he was a Postdoctoral Research Fellow with the Department of Automation, Tsinghua University. He has published more than 100 scientific papers in major journals and conferences. His current research interests include multimedia information retrieval, computer vision, and machine learning. He served as a Leading Guest

Editor for NPL and MTAP; the Special Session Chair for ICASSP 2021, ICME 2020/2019, and PCM 2017; and a reviewer for more than 20 prestigious international journals and conferences.



Guoping Liu (Member, IEEE) received the M.S. degree from Zhejiang University in 2016. He is currently an Expert Algorithm Engineer with Didi Chuxing. His research interests include spatial-temporal modeling and multi-modal fusion in intelligent transportation.