Contrastive Hypergraph Source-free Domain Adaptive Object Detection in Adverse Weather Conditions

Jicheng Yuan[*†]
Open Distributed Systems
Technical University of Berlin
jicheng.yuan@tu-berlin.de

Tung Kieu[†]
Department of Computer Science
Aalborg University, Denmark
tungkvt@cs.aau.dk

Manh Nguyen-Duc
Open Distributed Systems
Technical University of Berlin
duc.manh.nguyen@tu-berlin.de

Manfred Hauswirth
Open Distributed Systems
Technical University of Berlin
Fraunhofer FOKUS
manfred.hauswirth@tu-berlin.de

Danh Le-Phuoc
Open Distributed Systems
Technical University of Berlin
Fraunhofer FOKUS
danh.lephuoc@tu-berlin.de

Abstract

Adverse weather conditions such as fog, rain, snow, and low illumination introduce structured visibility degradations that severely impair object detectors trained on clear-weather data. When source data are inaccessible due to privacy or transmission constraints, source-free domain adaptation (SFDA) becomes necessary. However, existing SFDA methods based on pairwise contrastive learning struggle to capture high-order semantics and often suffer from agreement collapse under severe corruption. We propose REUNION, a hypergraph-guided SFDA framework for object detection in adverse weather conditions. REUNION constructs a contrastive hypergraph over target-domain object proposals, encoding high-order relations through intra-image context, weather-aware grouping, prototype-based semantic anchors, and uncertainty-aware connections between reliable and low-confidence instances. To effectively exploit these structured relations, we introduce a group-wise Hyper-InfoNCE objective that optimizes representations at the hyperedge level, enabling semantic information to propagate from confident proposals to corrupted or low-contrast instances. Experiments with diverse benchmarks demonstrate that REUNION effectively mitigates domain shifts and achieves state-of-the-art performance in SFDA settings. Code is available at this link.

1 Introduction

Reliable perception is a cornerstone of modern on-vehicle intelligent systems, supporting safety-critical applications such as autonomous driving and advanced driver-assistance systems. Object detection is a fundamental component of this pipeline, enabling downstream tasks, including tracking,

[*]Corresponding author.
[†]Equal contribution.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

motion prediction, and planning. Therefore, detection failures can propagate to higher-level decision modules and pose severe safety risks [30, 37]. Despite substantial progress under favorable conditions, reliable object detection under adverse weather conditions remains a persistent problem. Fog, rain, snow, and low-illumination scenes induce *structured* visibility degradations—such as contrast attenuation, color shifts, and partial occlusions—that systematically alter scene characteristics. Thus, detectors trained on clear-weather data degrade sharply when deployed in these conditions, with failures most pronounced on small or low-contrast objects such as pedestrians, riders, poles, and bicycles.

Domain adaptation [41] provides a principled way to mitigate such distribution shifts, yet many real-world deployments cannot retain or access the original source training data. Privacy constraints, licensing restrictions, and storage or transmission costs often render the source images unavailable once a model is deployed. This motivates *source-free domain adaptation* (SFDA) [54], where adaptation must proceed using only a source-trained detector and unlabeled target-domain images. Compared to standard unsupervised domain adaptation (UDA) [27], SFDA excludes explicit source–target alignment and makes adaptation highly dependent on pseudo-label quality, often leading to unstable training. Taking this into account and benefiting from advanced diffusion models [3], we incorporate layout-preserving weather variants synthesized using a frozen generative model, FLUX.1 KONTEXT [28], to improve the target-level structural consistency in various weather conditions.

Adverse weather conditions further amplify two failure modes that destabilize SFDA for detection. First, confidence-based filtering tends to *over-prune* true positives: Many valid objects receive low confidence under visibility corruption and are discarded before contributing to learning. Second, existing SFDA methods largely rely on *pairwise graph-based or instance-level representation learning objectives*, such as consistency regularization or contrastive alignment [49, 20]. Under severe corruption or strong augmentations, these objectives are prone to *agreement collapse*, forcing alignment between semantically mismatched samples and reinforcing confirmation bias.

We argue that SFDA under adverse weather should exploit *group-level* structure that remains informative even when individual appearances are unreliable. In driving scenes, object proposals exhibit stable higher-order regularities, including contextual co-occurrence within an image, shared weather-dependent statistics, and latent semantic neighborhoods formed by confident instances. `REUNION` operationalizes this insight by reformulating SFDA as *hypergraph-guided representation learning*, as illustrated in Fig. 1. Specifically, we construct a contrastive hypergraph over target-domain proposals, where hyperedges encode complementary high-order relations: (i) intra-image contexts preserving scene structures, (ii) weather-aware groupings aligning domain-specific statistics, (iii) prototype-based clusters anchoring semantics to achieve class-consistent clustering across diverse weather conditions, and (iv) uncertainty-aware connections linking low-confidence proposals to reliable neighbors.

We then introduce a group-wise *Hyper-InfoNCE* objective that propagates semantic information through hyperedges, enabling robust adaptation under severe visibility degradation. Beyond the hypergraph formulation, `REUNION` is implemented within a teacher-student adaptation framework, where a teacher detector provides pseudo labels to supervise the student model during source-free adaptation. On top of this backbone, `REUNION` incorporates several stabilizers tailored to adverse-weather adaptation. We refine noisy teacher predictions through a semantic-level hypergraph regularization module, which reduces cross-class leakage and mitigates confirmation bias in pseudo labels. To further strengthen cross-condition consis-

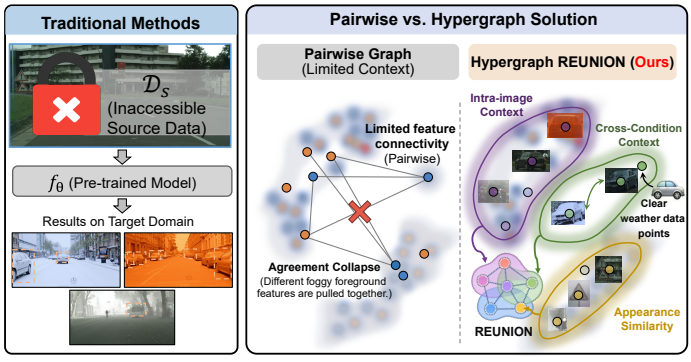


Figure 1: In SFDA, source data is inaccessible, while target images suffer from structured degradation (e.g., fog, rain, night), leading to detection collapse. Existing pairwise contrastive methods fail in such noisy settings. `REUNION` builds a contrastive hypergraph over object proposals using multi-relation hyperedges (e.g., intra-image context, and cross-condition views) to propagate semantics from confident to low-contrast instances.

tency, we synthesize layout-preserving weather variants using a frozen diffusion model, FLUX.1 KONTEXT [28], as a data augmentation strategy, avoiding generator training or access to labeled source data. Together, these components improve pseudo-label reliability and provide more robust supervision for learning under severe visibility degradation.

To verify our proposed framework, we conduct a comprehensive experiment on multiple adverse-weather benchmarking datasets. The experimental results demonstrate that `REUNION` consistently improves detection performance under severe visibility degradation and achieves state-of-the-art results in SFDA settings.

In summary, our contributions are: (i) We propose `REUNION`, a hypergraph-guided source-free domain adaptation framework for object detection under adverse weather. (ii) We construct a multi-relation contrastive hypergraph over target-domain object proposals and optimize it using a group-wise Hyper-InfoNCE objective, enabling semantic propagation through high-order relations and mitigating agreement collapse. (iii) We incorporate prototype-based semantic regularization and uncertainty-aware propagation to stabilize pseudo-label learning for low-contrast instances under severe visibility degradation. (iv) We demonstrate the effectiveness of `REUNION` through extensive experiments on multiple adverse-weather benchmarks under SFDA settings.

2 Related Work

2.1 Domain Adaptive Object Detection.

Domain adaptive object detection (DAOD) tackles the domain gap between a labeled source domain and an unlabeled target domain by aligning representations or predictions across domains [45]. Early and recent UDA-based DAOD methods adopt feature-level alignment, consistency regularization, and instance-aware objectives to mitigate distribution shift, yielding strong baselines such as MIC [16], MRT [59], 2PC [25], AT [32], and ALDI [24]. These approaches typically assume access to source data during adaptation. Furthermore, source-free domain adaptation (SFDA) focuses on addressing the challenges posed by domain discrepancies solely with pseudo-labels generated by a source-pretrained model while source data is inaccessible. Existing thresholding-based methods [33, 56, 31] rely solely on high-confidence predictions, discarding valuable information from uncertain candidates, yet potentially correct, detections. MixLRCo [57] propose low-confidence potential as pseudo-labels to maximize the utility of the target data. Beyond thresholding, other techniques have been proposed to refine pseudo-supervision. HCL [19] employs contrastive learning for improved instance and category discrimination, while SOAP [52] utilizes adversarial learning by injecting domain-specific noise into target images to regularize the detector. Upon this, to mitigate the domain shift, DiaNA [18] and DaC [58] divide the target domain into source-similar and source-dissimilar subsets, employing adversarial learning to align the feature representations of both subsets, thereby enhancing the quality of the pseudo-supervision. Similarly, LODS [31] aims to improve robustness by learning representations that are invariant to domain-specific styles, directly tackling the negative effects of domain shift rather than just filtering its symptoms (i.e., noisy labels). Additionally, to conquer training instability and enhance model robustness, PETS [34] and DDT [15] mitigate it separately by introducing multiple teachers and dynamic updating strategies, fostering a more stable and robust adaptation process.

2.2 Graph-based Approaches on Vision

Graph-based approaches have been widely used in computer vision to model relational structure among visual entities such as regions or objects. Early works construct pairwise graphs to encode semantic and spatial relations for scene understanding and visual reasoning, most notably in scene graph generation and object-centric reasoning [53, 36, 17]. Such graph formulations have also been extended to video understanding by connecting spatial regions across time for spatio-temporal reasoning [50]. However, these methods rely on pairwise edges and are limited in representing higher-order group structure in complex scenes. To overcome this limitation, recent works explore hypergraph formulations that model high-order relations via multi-node hyperedges. Representative approaches include Vision HGNN [13], HG-SFDA [22] and HG-ViT [11], which integrate hypergraph computation into recognition architectures, as well as hypergraph-based detection frameworks such as Hyper-YOLO [10]. These methods demonstrate the effectiveness of hypergraphs for visual representation learning, but typically assume fully supervised training and use hypergraphs as forward

computational modules. In contrast, our work leverages hypergraphs in an SFDA setting for object detection under adverse weather, constructing both proposal-level and semantic-level hypergraphs on unlabeled target data and using them to guide contrastive learning for robust semantic propagation.

3 Methodology

3.1 Problem Settings

Let $\mathcal{X} = \{\mathbf{I} \in \mathbb{R}^{H\times W\times 3}\}$ denote the image space, where $\mathbf{I}$ represents an RGB image with height H and width W. Let $\mathcal{Y} = \{\mathbf{B}, \mathbf{c}, \mathbf{T}\}$ denote the detection output space, where $\mathbf{B}$ denotes bounding box coordinates, $\mathbf{c}$ denotes object category labels, and $\mathbf{T}$ denotes bounding box regression targets.

We consider a labeled source-domain detection dataset, $\mathcal{D}_s = \left\{(\mathbf{I}_i^s, \{\mathbf{B}_{i,r}^s, \mathbf{c}_{i,r}^s, \mathbf{T}_{i,r}^s\}_r)\right\}_{i=1}^{N_s}$, where $\mathbf{I}_i^s \in \mathcal{X}$ is a source image and $\{\mathbf{B}_{i,r}^s, \mathbf{c}_{i,r}^s, \mathbf{T}_{i,r}^s\}_r$ denotes the set of proposal-level training instances in image i. Here, $\mathbf{B}_{i,r}^s$ is the region proposal of the r-th object, $\mathbf{c}_{i,r}^s \in \{1,\dots,C\}$ is its class label, and $\mathbf{T}_{i,r}^s$ is the corresponding regression target.

A detector F_{θ_0}, parameterized by θ_0, is trained on $\mathcal{D}_s$ using a standard supervised detection objective:

$$\mathcal{L}_{\text{det}}^s = \frac{1}{N_s}\sum_{i=1}^{N_s}\sum_r \Big[\mathcal{L}_{\text{cls}}(f_{\theta_0}(\mathbf{B}_{i,r}^s), \mathbf{c}_{i,r}^s) + \lambda\mathcal{L}_{\text{reg}}(r_{\theta_0}(\mathbf{B}_{i,r}^s), \mathbf{T}_{i,r}^s)\Big], \tag{1}$$

where f_{θ_0} and r_{θ_0} denote the classification and regression heads, respectively, and λ balances the two loss terms.

After training, the source dataset $\mathcal{D}_s$ is no longer accessible due to privacy, storage, or transmission constraints. Instead, we are given an unlabeled target-domain dataset $\mathcal{D}_t = \{(\mathbf{I}_i^t, w_i)\}_{i=1}^{N_t}$, where $\mathbf{I}_i^t$ is a target image and $w_i \in \{1,\dots,W\}$ is a discrete weather label (e.g., clear, rainy, foggy, snowy). These weather labels are obtained from external sources such as metadata or a frozen, pretrained weather classifier [12], and do not require access to source-domain data.

Inspired by recent advances in the controllability of modern text-to-image (T2I) diffusion models, we use FLUX.1 KONTEXT [28] to synthesize layout-preserving cross-condition views for target images, producing semantically consistent variations across weather conditions while preserving scene structure. These generated images augment the target-domain set $\mathcal{D}_t$ and provide additional supervision for adaptation.

Due to adverse weather conditions, the source and target domains follow different joint distributions, i.e., $P_s(X,Y) \neq P_t(X,Y)$, where the shift is primarily caused by structured visibility degradation, including reduced contrast, color distortion, and occlusion [21]. Our objective is to adapt the pretrained detector from θ_0 to new parameters θ using only unlabeled target data, such that the detection performance on the target domain is maximized.

3.2 Framework Overview

We adopt a mean-teacher framework [45], which consists of two detectors with identical architectures: a student model F_θ and a teacher model F_ϕ. The student model is optimized by gradient descent [9], while the teacher model is updated as an exponential moving average (EMA) [29] of the student via $\phi \leftarrow \alpha\phi + (1-\alpha)\theta$, where $\alpha \in (0,1)$ controls the temporal smoothness of the teacher. This formulation stabilizes training and provides more reliable pseudo labels for self-training. Since source data is inaccessible during adaptation, all region proposals, embeddings, and graph structures in the following are constructed exclusively from target-domain images; for notation simplicity, we omit the superscript ^{t} unless ambiguity may arise.

Given a target image $\mathbf{I}_i$, the detector generates a set of region proposals $\mathcal{B}_i = \{\mathbf{B}_{i,r}\}_{r=1}^{R_i}$ using a region proposal network (RPN) [42]. For each proposal $\mathbf{B}_{i,r}$, the teacher model outputs a class probability distribution $f_\phi(\mathbf{B}_{i,r})$ and a regression prediction $r_\phi(\mathbf{B}_{i,r})$.

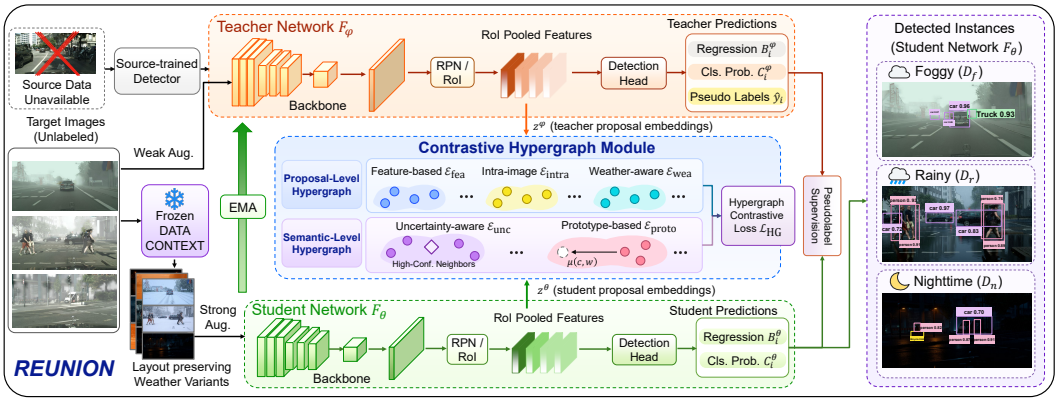


Figure 2: Overview of REUNION. Under the source-free setting, we employ a Teacher-Student framework to adapt to adverse weather conditions. The core component is the Contrastive Hypergraph Module (center), which aggregates object proposals into contrastive groups via intra-image, weather-aware, prototype-based, and uncertainty-aware hyperedges. These structural constraints guide the student's representation learning through a Hyper-InfoNCE loss, effectively aligning features across diverse weather conditions and reuniting fragmented semantics.

3.3 Pseudo Labeling

Pseudo labels are constructed from the teacher predictions as $\hat{\mathbf{c}}_{i,r} = \arg\max_c f_\phi(\mathbf{B}_{i,r})_c, \quad \hat{\mathbf{T}}_{i,r} = r_\phi(\mathbf{B}_{i,r})$. Using these pseudo labels, the student model is trained with an unsupervised detection loss:

$$\mathcal{L}_{\text{det}} = \frac{1}{N_t} \sum_{i=1}^{N_t} \sum_r \Big[\mathcal{L}_{\text{cls}}(f_\theta(\mathbf{B}_{i,r}), \hat{\mathbf{c}}_{i,r}) + \lambda \mathcal{L}_{\text{reg}}(r_\theta(\mathbf{B}_{i,r}), \hat{\mathbf{T}}_{i,r}) \Big]. \tag{2}$$

To further enforce consistency between the student and teacher predictions, we introduce a knowledge distillation loss on the classification logits:

$$\mathcal{L}_{\text{KD}} = \frac{1}{N_t} \sum_{i=1}^{N_t} \sum_r D_{\text{KL}} \left[\sigma\left(\frac{f_\phi(\mathbf{B}_{i,r})}{\tau_{\text{KD}}} \right) \middle\| \sigma\left(\frac{f_\theta(\mathbf{B}_{i,r})}{\tau_{\text{KD}}} \right) \right], \tag{3}$$

where $D_{\text{KL}}(\cdot)$ denotes the Kullback-Leibler divergence, $\sigma(\cdot)$ denotes the *softmax* function, and $f_\phi(\mathbf{B}_{i,r})$ and $f_\theta(\mathbf{B}_{i,r})$ represent the output classification logits of the teacher and student models, respectively. τ_{KD} is a temperature hyperparameter.

3.4 Proposal-Level Hypergraph

To capture higher-order structural relations among target proposals, we model proposals within a mini-batch as a weighted hypergraph. Specifically, we flatten all region proposals in a mini-batch and index them using a single proposal index $p \in \{1, \ldots, N_b\}$, where N_b denotes the total number of proposals in the mini-batch. Each proposal $\mathbf{B}_p$ is treated as a vertex v_p. We extract proposal-level embeddings from the teacher model and normalize them via $\mathbf{z}_i = \dfrac{g_\phi^{\text{emb}}(\mathbf{B}_p)}{\|g_\phi^{\text{emb}}(\mathbf{B}_p)\|_2}$,: where $g_\phi^{\text{emb}}(\cdot)$ denotes the embedding head.

We construct three types of hyperedges. First, feature-based hyperedges $\mathcal{E}_{\text{fea}}$ connect each proposal to its K nearest neighbors in the embedding space, capturing local semantic similarity. Second, intra-image hyperedges $\mathcal{E}_{\text{img}}$ connect all proposals originating from the same target image, encoding scene-level spatial and contextual coherence. Third, weather-aware hyperedges $\mathcal{E}_{\text{wea}}$ connect proposals extracted from images sharing the same weather label, encouraging weather-invariant representation learning.

The full proposal-level hyperedge set is $\mathcal{E}^{\text{prop}} = \mathcal{E}_{\text{fea}} \cup \mathcal{E}_{\text{img}} \cup \mathcal{E}_{\text{wea}}$. Together with the semantic-level hyperedges introduced below, the complete hyperedge set is $\mathcal{E} = \mathcal{E}^{\text{prop}} \cup \mathcal{E}_{\text{proto}} \cup \mathcal{E}_{\text{unc}}$. Let $H \in$

$\{0,1\}^{N_b \times |\mathcal{E}|}$ denote the hypergraph incidence matrix, where $H_{p,e} = 1$ indicates that vertex v_p belongs to hyperedge e. The induced adjacency matrix is defined as $A_{p,q} = \sum_{e \in \mathcal{E}} \omega_e H_{p,e} H_{q,e}$, equivalently $A = H \operatorname{diag}(\boldsymbol{\omega}) H^\top$, where ω_e denotes the weight of hyperedge e. We assign unit weights to feature-based, intra-image, weather-aware, and prototype-based hyperedges, while uncertainty-aware hyperedges are weighted by the normalized teacher prediction entropy, as detailed in Section 3.5.

3.5 Semantic-Level Hypergraph

To provide global semantic anchors and mitigate pseudo-label drift, we maintain class- and weather-conditioned prototypes in the embedding space. For each class $c \in \{1, \ldots, C\}$ and weather type $w \in \{1, \ldots, W\}$, we maintain a prototype vector $\mu_{c,w}$. At each iteration, the prototype is updated using an EMA, $\mu_{c,w} \leftarrow (1-\gamma)\mu_{c,w} + \gamma \frac{1}{|\mathcal{S}_{c,w}|} \sum_{p \in \mathcal{S}_{c,w}} \mathbf{z}_p$, where $\mathcal{S}_{c,w}$ denotes the set of high-confidence proposals in the mini-batch assigned to class c under weather w. Proposals that are close to the same prototype $\mu_{c,w}$ in embedding space are grouped to form prototype-based hyperedges denoted as $\mathcal{E}_{\text{proto}}$, which encourage class-consistent clustering across weather conditions. Specifically, we define that proposal v_p belongs to $\mathcal{E}_{\text{proto}}$ if its cosine similarity with the prototype exceeds a predefined threshold τ_{proto}, as $\frac{\mathbf{z}_p^\top \mu_{c,w}}{\|\mathbf{z}_p\|_2 \|\mu_{c,w}\|_2} \geq \tau_{\text{proto}}$.

Adverse weather often causes true object instances to receive low-confidence predictions. Instead of discarding these proposals, we explicitly model their relationships with reliable ones using uncertainty-aware hyperedges.

Teacher uncertainty is quantified using normalized entropy, $\mathcal{H}_p = -\frac{1}{\log C} \sum_c f_\phi(\mathbf{B}_p)_c \log f_\phi(\mathbf{B}_p)_c$. Low-confidence proposals are connected to their K nearest high-confidence neighbors, which is set by default at ten, that share the same pseudo class and weather condition. These connections form uncertainty-aware hyperedges denoted as $\mathcal{E}_{\text{unc}}$, whose weights are set proportional to $\mathcal{H}_p$. This design enables semantic information to propagate from reliable proposals to uncertain ones, improving robustness under severe visibility degradation. In addition, with the constructed hypergraph, let $Z \in \mathbb{R}^{N_b \times d}$ denote the proposal embedding matrix within a mini-batch, where the p-th row corresponds to z_p. Using the induced adjacency matrix A composed of all hyperedges ($\mathcal{E}_{fea}, \mathcal{E}_{img}, \mathcal{E}_{wea}, \mathcal{E}_{proto}, \mathcal{E}_{unc}$), we perform a graph convolution step to aggregate features from structured neighbors, as $Z^{(l+1)} = \sigma(\tilde{D}^{-1/2}\tilde{A}\tilde{D}^{-1/2}Z^{(l)}W^{(l)})$, where $\tilde{A} = A + I$ is the adjacency matrix with self-loops, $\tilde{D}$ is the corresponding degree matrix ($\tilde{D}_{p,p} = \sum_q \tilde{A}_{p,q}$), $W^{(l)}$ is a learnable weight matrix. Through this forward message passing, isolated or low-confidence proposals dynamically absorb contextual and prototype-based semantics from reliable neighbors. The updated representations $Z^{(l+1)}$ are then utilized for the final detection predictions and the contrastive learning objective. In this way, the constructed HGNN acts as a relational regularizer before computing the contrastive loss.

3.6 Hypergraph-Guided InfoNCE Learning Objective

While the unsupervised detection and consistency losses enable self-training, they rely on individual pseudo labels and thus remain sensitive to noise introduced by adverse weather. To further exploit the structural relations encoded in the proposal-level hypergraph, we introduce a hypergraph-guided contrastive learning objective that enforces representation consistency at the group level.

For each proposal $\mathbf{B}_p$, we extract ℓ_2-normalized embeddings from both the student and teacher models: $\mathbf{z}_p^S = \frac{g_\theta^{\text{emb}}(\mathbf{B}_p)}{\|g_\theta^{\text{emb}}(\mathbf{B}_p)\|_2}$, $\mathbf{z}_p^T = \frac{g_\phi^{\text{emb}}(\mathbf{B}_p)}{\|g_\phi^{\text{emb}}(\mathbf{B}_p)\|_2}$, where g_θ^{emb} and g_ϕ^{emb} denote the embedding heads of the student and teacher, respectively. We measure similarity using temperature-scaled cosine similarity via $\text{sim}(\mathbf{z}_p^S, \mathbf{z}_q^T) = \frac{(\mathbf{z}_p^S)^\top \mathbf{z}_q^T}{\tau}$, where τ is a temperature hyperparameter.

For a given anchor proposal p, we define its positive set based on the hypergraph adjacency matrix $\mathbf{A}$. Specifically, proposal q is considered a positive of p if (i) $p \neq q$, (ii) $\mathbf{A}_{p,q} > 0$, and (iii) both

proposals share the same pseudo class label. Formally, the positive set is defined as $\mathcal{P}_p = \{q \mid q \neq p,\ \mathbf{A}_{p,q} > 0,\ \hat{\mathbf{c}}_q = \hat{\mathbf{c}}_p\}$. This definition ensures that only structurally related and semantically compatible proposals are pulled together in the embedding space.

Different hyperedges encode relations of varying reliability. To account for this, we weight each positive pair (p, q) by the normalized hypergraph adjacency: $\tilde{w}_{p,q} = \dfrac{\mathbf{A}_{p,q}}{\sum_{k \in \mathcal{P}_p} \mathbf{A}_{p,k} + \epsilon}$, where ϵ is a small constant for numerical stability.

The hypergraph-guided InfoNCE loss for anchor p is then defined as

$$\mathcal{L}_{\text{HG-NCE}} = -\sum_{q \in \mathcal{P}_p} \tilde{w}_{p,q} \log \frac{\exp(\text{sim}(\mathbf{z}_p^S, \mathbf{z}_q^T))}{\sum_{k \neq p} \exp(\text{sim}(\mathbf{z}_p^S, \mathbf{z}_k^T))}. \tag{4}$$

The denominator includes all proposals in the same mini-batch, encouraging the student embedding to be close to hypergraph-consistent positives while being separated from unrelated proposals.

Overall Hypergraph Contrastive Loss. The final hypergraph-guided InfoNCE objective is obtained by averaging over all anchors with non-empty positive sets:

$$\mathcal{L}_{\text{HG}} = \frac{1}{|\mathcal{I}|} \sum_{p \in \mathcal{I}} \mathcal{L}_{\text{HG-NCE}}(p), \quad \mathcal{I} = \{p \mid |\mathcal{P}_p| > 0\}. \tag{5}$$

This objective enforces consistency between student and teacher representations at the level of hypergraph-defined groups, allowing semantic information to propagate across intra-image context, weather conditions, prototype anchors, and uncertainty-aware relations. As a result, low-confidence proposals can recover meaningful representations by leveraging their connections to reliable neighbors, which is critical under severe visual degradation.

4 Experiments

4.1 Experimental Settings

Datasets and Scenarios. The experiments are conducted on three datasets: Cityscapes [7], Cityscapes-Adverse, and Sim10k [23]. We conduct experiments on two scenarios: (1) City scene source-free adaptation: Cityscapes $\rightarrow$ Cityscapes-Adverse; (2) Sim2Real source-free adaptation: Sim10k $\rightarrow$ Cityscapes-Adverse.

Baselines. We compare `REUNION` against representative object detection methods under domain shift, covering source-only, fully supervised, UDA, and SFDA settings. The Source-Only model, trained only on Cityscapes and directly evaluated on Cityscapes-Adverse, serves as the lower bound, while the Oracle model, trained with labeled target-domain data, provides an upper-bound reference. We also include several UDA methods, including MIC [16], MRT [59], 2PC [25], AT [32], and ALDI [24]. These methods assume access to both source images and unlabeled target images during adaptation, and therefore serve as reference points rather than direct competitors. Our primary comparisons are with SFDA methods, including LPLD [55], MT [45], DRU [26], IRG [49], and SF-UT [14], which follow the same source-free constraint as `REUNION` by adapting a source-pretrained detector using only unlabeled target-domain data.

Metrics. We evaluate object detection performance using two widely adopted metrics: Intersection over Union (IoU) and Average Precision (AP). We report AP computed at an IoU threshold of 0.5, following standard evaluation protocols in object detection.

Implementation and Hyperparameter Details. We adopt R-CNN with pre-trained ResNet-50 as the primary object detector. During the learning phase, the student network is initiated with a learning rate of 0.04, and we set the EMA for updating the teacher model to 0.9996. We set $\lambda = 1.0$ (see Equations 1 and 2) by default. All the experiments are conducted on four Nvidia A40 GPUs, with each GPU handling a batch size of 16. The implementation is carried out using the PyTorch framework and Detectron2 [51].

Table 1: Performance comparison of UDA/SFDA methods on Cityscapes → Cityscapes-Adverse with ResNet-50 backbones. All adaptation methods use the same offline FLUX-generated weather variants while preserving their method-specific online augmentations.

Type	Method	Detector	D-Rainy	D-Snowy	D-Foggy	D-Sandstorm	N-Clear	N-Rainy	N-Foggy	mAP
Lower-Bound	Source-Only	FRCNN-R50	15.9	16.1	20.8	3.0	9.2	6.4	5.5	11.0
Higher-Bound	Oracle	FRCNN-R50	41.9	50.9	48.1	45.7	39.7	35.5	37.1	42.7
UDA	MIC [16]	FRCNN-R50	21.0	23.2	29.9	26.2	17.3	10.9	13.8	20.3
	MRT [59]	Def-DETR-R50	18.4	25.2	29.6	18.3	17.3	7.0	14.6	18.6
	2PC [25]	FRCNN-R50	37.4	31.3	42.5	17.0	27.9	28.6	22.1	29.5
	AT [32]	FRCNN-R50	32.1	26.4	41.7	35.5	27.6	27.4	19.7	30.1
	ALDI [24]	FRCNN-R50	39.8	36.0	54.1	44.0	35.5	27.9	26.6	37.7
SFDA	DRU [26]	Def-DETR-R50	32.1	30.6	42.4	32.3	27.1	16.5	20.7	28.8
	SF-UT [14]	FRCNN-R50	30.8	33.1	48.0	37.9	29.0	14.5	23.6	31.0
	IRG [49]	FRCNN-R50	35.0	33.0	33.8	34.4	30.4	**31.8**	27.6	32.3
	DDT [15]	Def-DETR-R50	33.5	30.2	40.8	33.9	25.6	19.3	24.1	29.6
	FALCON [48]	FRCNN-R50	36.7	35.8	46.9	39.3	33.1	23.3	29.1	34.9
	CGSA [8]	RT-DETR-R50	40.4	35.8	48.5	44.1	32.9	30.0	25.5	36.7
	REUNION	FRCNN-R50	43.7	38.2	**53.9**	**45.7**	36.1	26.6	30.8	39.3
	REUNION	Def-DETR-R50	**45.5**	**39.6**	52.8	45.2	**36.3**	28.4	**32.5**	**40.0**

4.2 Experimental Findings

As shown in Table 1, we evaluate REUNION in the setting Cityscapes → Cityscapes-Adverse against leading methods in both source-free domain adaptation (SFDA) and unsupervised domain adaptation (UDA). The proposed method achieves the best overall performance, improving AP by approximately 3.3 points over the strongest SFDA baseline, CGSA [8], while remaining competitive with several UDA approaches that have access to source data.

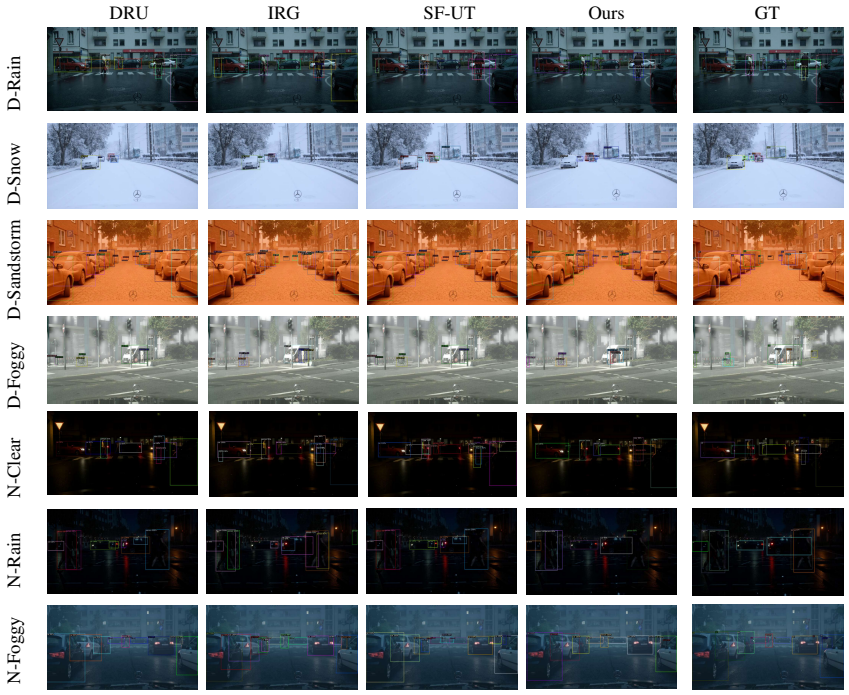


Figure 3: Qualitative results for the Cityscapes-Adverse dataset are displayed from left to right, showcasing the results from DRU [26], IRG [49], SF-UT [14], and the proposed method REUNION, with the ground truth presented on the far right.

Notably, REUNION performs particularly well under challenging conditions such as rainy, sandstorm, and night-foggy. Qualitative results in Fig. 3 further illustrate that REUNION produces more

accurate detections across diverse illumination and weather conditions, including low-visibility scenes with artificial lighting. In contrast, competing methods such as IRG [49] and SF-UT [14] often produce false positives in dark regions or miss objects under severe illumination changes.

These results indicate that the proposed hypergraph-guided learning framework improves representation robustness across adverse weather conditions and varying object scales. We further evaluate REUNION in a challenging Sim2Real setting by adapting a detector trained on the synthetic Sim10k dataset to the real-world Cityscapes-Adverse benchmark. As shown in Table 2, REUNION achieves the best performance among all SFDA methods, improving AP by 1.8% over the strongest baseline, IRG [49]. Notably, our method also performs competitively with UDA approaches that rely on source data during adaptation. This setting introduces substantial domain discrepancies, including differences in rendering realism, scene composition, and weather-induced visibility degradation. Despite these challenges, REUNION consistently improves detection accuracy, demonstrating that hypergraph-guided representation learning effectively facilitates semantic transfer across both synthetic-to-real and adverse-weather domain shifts.

Table 2: Sim2Real Adaptation from Sim10K → Cityscapes-Adverse.

Type	Method	AP Car
S	Source Only	19.5
UDA	DAF [4]	22.4
	PT [2]	22.7
	MIC [16]	27.5
	AT [32]	28.4
	2PC [25]	31.2
	ALDI [24]	42.4
SFDA	LPLD [55]	32.0
	MT [45]	33.1
	DRU [26]	37.0
	SF-UT [14]	39.7
	IRG [49]	40.8
	REUNION (**ours**)	42.6

We further compare REUNION with SFDA baselines on Cityscapes → Cityscapes-Adverse scenarios according to different APs for different object sizes. As shown in Table 3, REUNION achieves the best performance on large and small objects, reaching 34.6 AP^L and 8.4 AP^S. These gains indicate that the proposed hypergraph formulation effectively propagates semantic information across related proposals, helping maintain reliable representations despite weather-induced degradation. Improvements are particularly evident for large objects, where contextual cues are more informative. While gains on small objects remain limited due to extremely low visibility, the results demonstrate that hypergraph-guided learning improves detection robustness under adverse conditions.

4.3 Ablation Study

To assess the contribution of each component in REUNION, we performed ablation experiments on the Cityscapes → Cityscapes-Adverse benchmark in different configurations. As shown in Table 4, the baseline model using only weak-strong augmentation achieves limited performance. Introducing pairwise augmentation improves the performance to 34.1% AP. However, relying solely on pairwise contrastive views is insufficient to bridge the large domain gap caused by structured visibility degradation, often leading to agreement collapse.

By incorporating the proposal-level and semantic-level hyperedges, REUNION achieves 38.7% AP even without the Hyper-InfoNCE objective, indicating that modeling higher-order relations among proposals provides more reliable semantic cues for learning under visual corruption. This is attributed to the forward message passing mechanism defined in REUNION. Specifically, the HGNN serves as a structural prior and provides relational regularization that guides the student model's representation learning, allowing confident instances to calibrate the features of corrupted ones (e.g., via uncertainty-aware edges) before they are passed to the detection head. Thus, the HGNN prevents the student model from collapsing under severe weather degradation by achieving relational consistency at the feature level. To further analyze the contribution of different hyperedge types, we perform incremental experiments as shown in Table 5. Starting with feature-based hyperedges ($\mathcal{E}_{\text{fea}}$), the model captures local semantic similarity among proposals. Adding intra-image ($\mathcal{E}_{\text{img}}$) and weather-aware ($\mathcal{E}_{\text{wea}}$) hyperedges improves performance by 0.7% AP, demonstrating the benefit of incorporating scene-level context and domain-specific grouping. Further introducing prototype-based ($\mathcal{E}_{\text{proto}}$) and uncertainty-aware ($\mathcal{E}_{\text{unc}}$) hyperedges yields the best performance of 39.3% AP on Cityscapes-Adverse. These hyperedges anchor proposal features to stable teacher prototypes and connect low-confidence proposals to reliable neighbors, enabling the model to recover informative instances that would

Table 3: Comparison to SFDA baselines with different object sizes on Cityscapes-Adverse.

Type	Method	AP^S	AP^M	AP^L
SFDA	LPLD [55]	5.2	14.3	19.4
	MT [45]	5.3	15.6	19.8
	IRG [49]	7.4	12.4	29.4
	SF-UT [14]	6.7	11.8	28.0
	REUNION (**ours**)	8.4	11.1	34.6

otherwise be discarded under severe visibility degradation. Overall, integrating all hyperedge families enables REUNION to better bridge the domain gap between clear and adverse weather conditions.

Table 4: Ablation study on Cityscapes-Adverse.

PairAug	Proposal HGNN	Semantic HGNN	Hyper-Info-NCE	AP	AP^L	AP^S	AP^M
				30.0	28.6	6.9	10.7
✓				34.1	30.7	7.5	10.8
✓		✓		37.1	33.0	7.9	11.0
✓	✓			37.9	33.7	8.1	11.6
✓	✓	✓		38.7	34.1	8.2	11.5
✓	✓	✓	✓	39.3	34.6	8.4	11.1

Table 5: Impact of hyperedge types in Hypergraph-based SFDA.

Hyperedge Types ($\mathcal{E}^{(t)}$)					AP	AP^L	AP^S	AP^M
Fea	Img	Weath	Proto	Unc				
✓					37.2	33.0	7.9	11.1
✓	✓				37.4	33.2	8.1	11.6
✓	✓	✓			37.9	33.7	8.1	11.6
✓	✓	✓	✓		38.8	33.7	7.9	9.3
✓	✓	✓	✓	✓	39.3	34.6	8.4	11.1

In addition, Fig. 4 shows the class-wise AP across different weather conditions. REUNION consistently outperforms existing SFDA methods across most categories, with larger gains under challenging conditions such as rain, snow, sandstorm, and night-foggy. This improvement highlights the benefit of modeling higher-order relations among proposals. Unlike pairwise methods, the proposed hypergraph enables semantic propagation among related instances, allowing the model to recover reliable representations for corrupted or low-confidence proposals and maintain stable detection performance across diverse weather conditions.

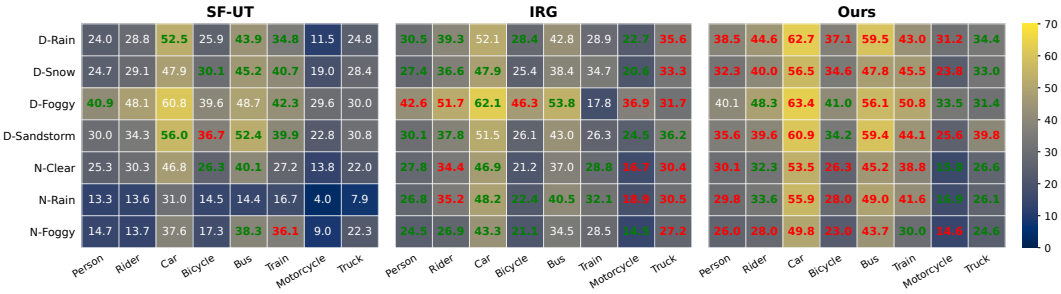


Figure 4: Class-wise AP across diverse weather conditions and models, IRG [49], SF-UT [14], and REUNION; best and second-best results are highlighted in red and green, respectively.

5 Conclusion and Future Work

We presented REUNION, a source-free domain adaptation framework for object detection under adverse weather. REUNION reformulates SFDA as hypergraph-guided representation learning, organizing proposal relations with intra-image context, weather-aware grouping, prototype anchors, and uncertainty-aware connections. The group-wise Hyper-InfoNCE objective propagates semantics from reliable to low-confidence instances and mitigates agreement collapse under severe visibility degradation. Experiments on Cityscapes-Adverse and cross-domain settings demonstrate consistent gains over strong UDA and SFDA baselines, validating the value of higher-order relational constraints in source-free adaptation. Future work includes extending REUNION to broader adverse conditions (mixed weather and low-light), improving pseudo-label calibration under severe corruption, and developing lighter hypergraph message-passing schemes for better efficiency.

6 Acknowledgements

This work is supported by the Deutsche Forschungsgemeinschaft, German Research Foundation under grant number 453130567 (COSMO), by the Horizon Europe Research and Innovation Actions under grant number 101092908 (SmartEdge), by the Chips Joint Undertaking (JU), European Union (EU) HORIZON-JU-IA, under grant agreement No. 101140087 (SMARTY), by the Federal Ministry for Education and Research, Germany under grant number BIFOLD25B.

Competing Interests. The authors declare no competing financial interests related to this work.

References

[1] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Ming-Hsuan Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report. *CoRR*, abs/2502.13923, 2025.

[2] Meilin Chen, Weijie Chen, Shicai Yang, Jie Song, Xinchao Wang, Lei Zhang, Yunfeng Yan, Donglian Qi, Yueting Zhuang, Di Xie, and Shiliang Pu. Learning domain adaptive object detection with probabilistic teacher. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 3040–3055, 2022.

[3] Minshuo Chen, Song Mei, Jianqing Fan, and Mengdi Wang. An overview of diffusion models: Applications, guided generation, statistical rates and optimization. *CoRR*, abs/2404.07771, 2024.

[4] Yuhua Chen, Wen Li, Christos Sakaridis, Dengxin Dai, and Luc Van Gool. Domain adaptive faster R-CNN for object detection in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3339–3348, 2018.

[5] Yuhua Chen, Haoran Wang, Wen Li, Christos Sakaridis, Dengxin Dai, and Luc Van Gool. Scale-aware domain adaptive faster R-CNN. *Int. J. Comput. Vis.*, 129(7):2223–2243, 2021.

[6] Ting-Rui Chiang, Joshua Robinson, Xinyan Velocity Yu, and Dani Yogatama. Locatebench: Evaluating the locating ability of vision language models. *CoRR*, abs/2410.19808, 2024.

[7] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. *CoRR*, abs/1604.01685, 2016.

[8] Boyang Dai, Zeng Fan, Zihao Qi, Meng Lou, and Yizhou Yu. CGSA: class-guided slot-aware adaptation for source-free object detection. In Carl Vondrick, Bharath Hariharan, Colin Raffel, Lerrel Pinto, Diyi Yang, and Aleksandra Faust, editors, *The Fourteenth International Conference on Learning Representations, ICLR 2026, Rio de Janeiro, Brazil, April 23-27, 2026*. proceedings.iclr.cc, 2026.

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

[10] Yifan Feng, Jiangang Huang, Shaoyi Du, Shihui Ying, Jun-Hai Yong, Yipeng Li, Guiguang Ding, Rongrong Ji, and Yue Gao. Hyper-yolo: When visual object detection meets hypergraph computation. *IEEE Trans. Pattern Anal. Mach. Intell.*, 47(4):2388–2401, 2025.

[11] Joshua Fixelle. Hypergraph vision transformers: Images are more than nodes, more than edges. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9751–9761, 2025.

[12] Jose Carlos Villarreal Guerra, Zeba Khanam, Shoaib Ehsan, Rustam Stolkin, and Klaus McDonald-Maier. Weather classification: A new multi-class dataset, data augmentation approach and comprehensive evaluations of convolutional neural networks. In *2018 NASA/ESA Conference on Adaptive Hardware and Systems (AHS)*, pages 305–310. IEEE, 2018.

[13] Yan Han, Peihao Wang, Souvik Kundu, Ying Ding, and Zhangyang Wang. Vision HGNN: an image is more than a graph of nodes. In *Proceedings of the IEEE/CVF Conference on Computer Vision (ICCV)*, pages 19821–19831, 2023.

[14] Yan Hao, Florent Forest, and Olga Fink. Simplifying source-free domain adaptation for object detection: Effective self-training strategies and performance insights. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 196–213, 2024.

[15] Qi He, Xiao Wu, Jun-Yan He, and Shuai Li. Dual-rate dynamic teacher for source-free domain adaptive object detection. In *IEEE/CVF International Conference on Computer Vision, ICCV 2025, Honolulu, HI, USA, October 19-25, 2025*, pages 2067–2076. IEEE, 2025.

[16] Lukas Hoyer, Dengxin Dai, Haoran Wang, and Luc Van Gool. MIC: masked image consistency for context-enhanced domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11721–11732, 2023.

[17] Han Hu, Jiayuan Gu, Zheng Zhang, Jifeng Dai, and Yichen Wei. Relation networks for object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3588–3597, 2018.

[18] Duojun Huang, Jichang Li, Weikai Chen, Junshi Huang, Zhenhua Chai, and Guanbin Li. Divide and adapt: Active domain adaptation via customized learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7651–7660, 2023.

[19] Jiaxing Huang, Dayan Guan, Aoran Xiao, and Shijian Lu. Model adaptation: Historical contrastive learning for unsupervised domain adaptation without source data. In *Proceedings of the Annual Conference on Neural Information Processing Systems (NeurIPS)*, pages 3635–3649, 2021.

[20] Uiwon Hwang, Jonghyun Lee, Juhyeon Shin, and Sungroh Yoon. SF(DA)2: source-free domain adaptation through the lens of data augmentation. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2024.

[21] Minsik Jeon, Junwon Seo, and Jihong Min. DA-RAW: domain adaptive object detection for real-world adverse weather conditions. In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, pages 2013–2020, 2024.

[22] Jinkun Jiang, Qingxuan Lv, Yuezun Li, Yong Du, Junyu Dong, Sheng Chen, and Hui Yu. HG-SFDA: hypergraph learning meets source-free unsupervised domain adaptation. *IEEE Trans. Image Process.*, 34:7542–7557, 2025.

[23] Matthew Johnson-Roberson, Charles Barto, Rounak Mehta, Sharath Nittur Sridhar, and Ram Vasudevan. Driving in the matrix: Can virtual worlds replace human-generated annotations for real world tasks? *CoRR*, abs/1610.01983, 2016.

[24] Justin Kay, Timm Haucke, Suzanne Stathatos, Siqi Deng, Erik Young, Pietro Perona, Sara Beery, and Grant Van Horn. Align and distill: Unifying and improving domain adaptive object detection. *Trans. Mach. Learn. Res.*, 2025, 2025.

[25] Mikhail Kennerley, Jian-Gang Wang, Bharadwaj Veeravalli, and Robby T. Tan. 2pcnet: Two-phase consistency training for day-to-night unsupervised domain adaptive object detection. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, pages 11484–11493. IEEE, 2023.

[26] Trinh Le Ba Khanh, Huy-Hung Nguyen, Long Hoang Pham, Duong Nguyen-Ngoc Tran, and Jae Wook Jeon. Dynamic retraining-updating mean teacher for source-free object detection. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 328–344, 2024.

[27] Elyor Kodirov, Tao Xiang, Zhen-Yong Fu, and Shaogang Gong. Unsupervised domain adaptation for zero-shot learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision (ICCV)*, pages 2452–2460, 2015.

[28] Black Forest Labs, Stephen Batifol, Andreas Blattmann, Frederic Boesel, Saksham Consul, Cyril Diagne, Tim Dockhorn, Jack English, Zion English, Patrick Esser, Sumith Kulal, Kyle Lacey, Yam Levi, Cheng Li, Dominik Lorenz, Jonas Müller, Dustin Podell, Robin Rombach, Harry Saini, Axel Sauer, and Luke Smith. FLUX.1 kontext: Flow matching for in-context image generation and editing in latent space. *CoRR*, abs/2506.15742, 2025.

[29] Samuli Laine and Timo Aila. Temporal ensembling for semi-supervised learning. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2017.

[30] Gim Hee Lee, Friedrich Fraundorfer, and Marc Pollefeys. Motion estimation for self-driving cars with a generalized camera. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2746–2753, 2013.

[31] Shuaifeng Li, Mao Ye, Xiatian Zhu, Lihua Zhou, and Lin Xiong. Source-free object detection by learning to overlook domain style. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8004–8013, 2022.

[32] Yu-Jhe Li, Xiaoliang Dai, Chih-Yao Ma, Yen-Cheng Liu, Kan Chen, Bichen Wu, Zijian He, Kris Kitani, and Peter Vajda. Cross-domain adaptive teacher for object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7571–7580, 2022.

[33] Jian Liang, Dapeng Hu, and Jiashi Feng. Do we really need to access the source data? source hypothesis transfer for unsupervised domain adaptation. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 6028–6039, 2020.

[34] Qipeng Liu, Luojun Lin, Zhifeng Shen, and Zhifeng Yang. Periodically exchange teacher-student for source-free object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 6391–6401, 2023.

[35] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, and Lei Zhang. Grounding DINO: marrying DINO with grounded pre-training for open-set object detection. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 38–55, 2024.

[36] Cewu Lu, Ranjay Krishna, Michael S. Bernstein, and Li Fei-Fei. Visual relationship detection with language priors. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 852–869, 2016.

[37] Ana I. Maqueda, Antonio Loquercio, Guillermo Gallego, Narciso García, and Davide Scaramuzza. Event-based vision meets deep learning on steering prediction for self-driving cars. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5419–5427, 2018.

[38] Matthias Minderer, Alexey A. Gritsenko, and Neil Houlsby. Scaling open-vocabulary object detection. In *Proceedings of the Annual Conference on Neural Information Processing Systems (NeurIPS)*, 2023.

[39] OpenAI. Gpt-4o system card. *CoRR*, abs/2410.21276, 2024.

[40] OpenAI. Introducing GPT-4.1 in the API. `https://openai.com/index/gpt-4-1/`, April 2025. Official OpenAI announcement.

[41] Poojan Oza, Vishwanath A. Sindagi, Vibashan VS, and Vishal M. Patel. Unsupervised domain adaptation of object detectors: A survey. *IEEE Trans. Pattern Anal. Mach. Intell.*, 46(6):4018–4040, 2024.

[42] Shaoqing Ren, Kaiming He, Ross B. Girshick, and Jian Sun. Faster R-CNN: towards real-time object detection with region proposal networks. In *Proceedings of the Annual Conference on Neural Information Processing Systems (NeurIPS)*, pages 91–99, 2015.

[43] Christos Sakaridis, Haoran Wang, Ke Li, René Zurbrügg, Arpit Jadon, Wim Abbeloos, Daniel Olmeda Reino, Luc Van Gool, and Dengxin Dai. ACDC: the adverse conditions dataset with correspondences for robust semantic driving scene perception. *IEEE Trans. Pattern Anal. Mach. Intell.*, 48(3):2970–2988, 2026.

[44] Ranjan Sapkota and Manoj Karkee. Object detection with multimodal large vision-language models: An in-depth review. *Inf. Fusion*, 126:103575, 2026.

[45] Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In *Proceedings of the Annual Conference on Neural Information Processing Systems (NeurIPS)*, pages 1195–1204, 2017.

[46] Gemini Team. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *CoRR*, abs/2507.06261, 2025.

[47] Qwen Team. Qwen3-vl technical report. *CoRR*, abs/2511.21631, 2025.

[48] Sairam VCR, Rishabh Lalla, Aveen Dayal, Tejal Kulkarni, Anuj Lalla, Vineeth N Balasubramanian, and Muhammad Haris Khan. Foundation model priors enhance object focus in feature space for source-free object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 29431–29440, 2026.

[49] Vibashan VS, Poojan Oza, and Vishal M. Patel. Instance relation graph guided source-free domain adaptive object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3520–3530, 2023.

[50] Xiaolong Wang and Abhinav Gupta. Videos as space-time region graphs. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 413–431, 2018.

[51] Yuxin Wu, Alexander Kirillov, Francisco Massa, Wan-Yen Lo, and Ross Girshick. Detectron2. `https://github.com/facebookresearch/detectron2`, 2019.

[52] Lin Xiong, Mao Ye, Dan Zhang, Yan Gan, Xue Li, and Yingying Zhu. Source data-free domain adaptation of object detector through domain-specific perturbation. *Int. J. Intell. Syst.*, 36(8):3746–3766, 2021.

[53] Danfei Xu, Yuke Zhu, Christopher B. Choy, and Li Fei-Fei. Scene graph generation by iterative message passing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3097–3106, 2017.

[54] Shiqi Yang, Yaxing Wang, Joost van de Weijer, Luis Herranz, and Shangling Jui. Generalized source-free domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision (ICCV)*, pages 8958–8967, 2021.

[55] Ilhoon Yoon, Hyeongjun Kwon, Jin Kim, Junyoung Park, Hyunsung Jang, and Kwanghoon Sohn. Enhancing source-free domain adaptive object detection with low-confidence pseudo label distillation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, volume 15142, pages 337–353, 2024.

[56] Fuming You, Jingjing Li, Lei Zhu, Zhi Chen, and Zi Huang. Domain adaptive semantic segmentation without source data. In *Proceedings of the ACM Multimedia Conference (MM)*, pages 3293–3302, 2021.

[57] Yixin Zhang, Junjie Li, and Zilei Wang. Low-confidence samples matter for domain adaptation. *CoRR*, abs/2202.02802, 2022.

[58] Ziyi Zhang, Weikai Chen, Hui Cheng, Zhen Li, Siyuan Li, Liang Lin, and Guanbin Li. Divide and contrast: Source-free domain adaptation via adaptive contrastive learning. In *Proceedings of the Annual Conference on Neural Information Processing Systems (NeurIPS)*, 2022.

[59] Zijing Zhao, Sitong Wei, Qingchao Chen, Dehui Li, Yifan Yang, Yuxin Peng, and Yang Liu. Masked retraining teacher-student framework for domain adaptive object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 18993–19003, 2023.

A Supplementary Material

The supplementary material provides additional details to complement the main paper. In particular, we offer further insights into the experimental datasets used in our study and present additional ablation experiments to analyze the contribution of different components of REUNION. These results provide a deeper understanding of the proposed method and further support its effectiveness for source-free domain adaptation in adverse-weather object detection.

A.1 Datasets and Evaluation Metrics

We evaluate the proposed REUNION on the Cityscapes-Adverse benchmark for source-free domain adaptation (SFDA) in adverse-weather object detection. Our experiments involve both real-world and synthetic datasets, including Cityscapes [7], Sim10K [23], and Cityscapes-Adverse. The Cityscapes-Adverse dataset is constructed from Cityscapes by introducing seven types of adverse weather and lighting conditions. In total, it contains 24,150 images covering eight object categories. The datasets are divided into training, adaptation, and evaluation splits following standard protocols [49, 20].

Table A1: Dataset statistics.

Dataset	**No. Classes**	**No. Images**		
		Source	Adaptation	Evaluation
Cityscapes[7]	8	2,950	-	-
Sim10K[23]	1	9,975	-	-
Cityscapes $\rightarrow$ Cityscapes-Adverse	8	-	20,650	3,500
Sim10K $\rightarrow$ Cityscapes-Adverse	1	-	19,768	3,346

For the Cityscapes $\rightarrow$ Cityscapes-Adverse scenario, we train the source detector using 2,950 daytime images from Cityscapes. During adaptation, 20,650 unlabeled images from Cityscapes-Adverse are used, while evaluation is performed on 3,500 Cityscapes-Adverse images.

For the Sim10K $\rightarrow$ Cityscapes-Adverse scenario, the source model is trained using 9,975 images from Sim10K. During adaptation, 19,768 unlabeled images from Cityscapes-Adverse are used, and the evaluation set contains 3,346 images from the same dataset.

Importantly, no target-domain annotations are used during the adaptation stage in either scenario. Table A1 summarizes the statistics of the experimental datasets.

Next, we analyze the distribution of object classes in Cityscapes, as illustrated in Fig. A1. The class distribution in Cityscapes-Adverse follows a similar pattern, since each Cityscapes image is used to generate seven corresponding images under different light and adverse-weather conditions. In contrast, the Sim10K dataset contains only a single object category, namely *car*.

A.2 Baseline Details

To comprehensively evaluate the performance of REUNION, we compare it with a wide range of state-of-the-art object detection methods under domain shift. The baselines are grouped according to their adaptation settings and data requirements, as summarized in Table 1.

The Source-Only model is trained exclusively on the source domain, Cityscapes, and evaluated directly on the target domain, Cityscapes-Adverse, without any adaptation. This baseline quantifies the performance degradation caused by the domain shift and serves as a lower-bound reference. In contrast, the Oracle model is trained using labeled target-domain data and represents an upper-bound reference for adaptation performance.

We also compare with several unsupervised domain adaptation (UDA) methods, including MIC [16], MRT [59], 2PC [25], AT [32], and ALDI [24]. These methods operate under a more privileged setting than ours, since they require access to both source-domain images and unlabeled target-domain images during adaptation. Therefore, they are not direct competitors to our source-free setting, but they provide useful reference points for understanding the performance achievable when source data is available for alignment.

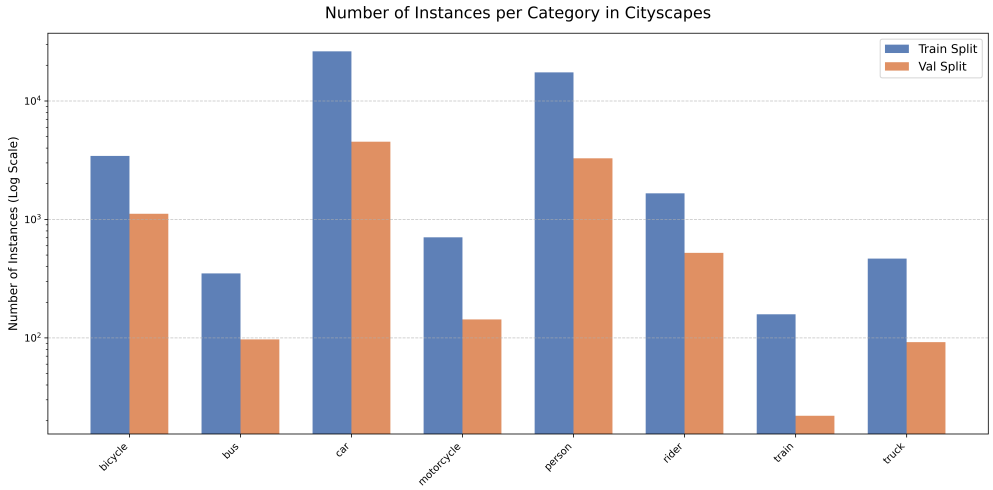


Figure A1: Class-wise instances distribution from Cityscapes dataset.

Table A2: Generalization across different detection architectures and backbones on Cityscapes-Adverse.

Method	Detector	D-Rainy	D-Snowy	D-Foggy	D-Sandstorm	N-Clear	N-Rainy	N-Foggy	mAP
DDT [15]	Def-DETR-R50	33.5	30.2	40.8	33.9	25.6	19.3	24.1	29.6
FALCON [48]	Faster R-CNN-R50	36.7	35.8	46.9	39.3	33.1	23.3	29.1	34.9
CGSA [8]	RT-DETR-R50	40.4	35.8	48.5	44.1	32.9	30.0	25.5	36.7
REUNION	Faster R-CNN-R50	43.7	38.2	53.9	45.7	36.1	26.6	30.8	39.3
REUNION	Def-DETR-R50	45.5	39.6	52.8	45.2	36.3	28.4	32.5	40.0
REUNION	Faster R-CNN-ViT-B	**48.0**	**44.3**	**57.2**	**50.3**	**39.6**	**31.4**	**37.7**	**44.1**

The most relevant comparisons are with source-free domain adaptation (SFDA) methods, including LPLD [55], MT [45], DRU [26], IRG [49], and SF-UT [14]. These methods follow the same challenging setting as `REUNION`: they adapt a source-pretrained detector using only unlabeled target-domain data, without access to the original source images. Thus, they serve as the direct competitors for evaluating the effectiveness of our proposed approach in the practical source-free setting.

A.3 Additional Experiments

Table A3 presents the detailed experimental results. The observations are consistent with those reported in the main paper. In particular, the proposed method achieves the best overall performance, improving AP by approximately 3.3 points over the strongest SFDA baseline, CGSA [8], while remaining competitive with several UDA approaches that have access to source data. Notably, `REUNION` performs especially well under challenging conditions such as rainy, sandstorm, and night-foggy.

A.4 Additional Ablation Study

A.4.1 Effect of Confidence Neighbors K.

To analyze the impact of the uncertainty-aware hyperedges ($\mathcal{E}_{\text{unc}}$), we evaluate the model under different values of K, which controls the number of high-confidence neighbors connected to each low-confidence proposal. As shown in Table A4, varying K influences how effectively semantic information propagates from reliable anchors to uncertain instances under severe visibility degradation.

When K is too small (e.g., $K = 6$ or $K = 8$), low-confidence proposals receive limited contextual support, making the learned representations sensitive to noisy predictions and local feature variations. As a result, detection performance across object scales (AP^L, AP^M, and AP^S) remains suboptimal.

Table A3: Detailed performance comparison of UDA and SFDA methods on Cityscapes → Cityscapes-Adverse across different detector architectures and backbones. All adaptation methods share the same offline FLUX-generated weather variants while retaining their method-specific online augmentations.

Type	Method	Detector	D-Rainy	D-Snowy	D-Foggy	D-Sandstorm	N-Clear	N-Rainy	N-Foggy	mAP
Lower-Bound	Source-Only	FRCNN-R50	15.9	16.1	20.8	3.0	9.2	6.4	5.5	11.0
Higher-Bound	Oracle	FRCNN-R50	41.9	50.9	48.1	45.7	39.7	35.5	37.1	42.7
UDA	DAF [4]	FRCNN-R50	25.0	23.2	32.8	30.0	18.6	11.1	15.3	22.3
	SADAF [5]	FRCNN-R50	21.6	22.9	33.7	27.4	17.0	9.0	15.2	20.1
	PT [2]	FRCNN-VGG16	23.6	21.5	35.9	28.3	18.6	16.0	14.7	22.7
	MIC [16]	FRCNN-R50	21.0	23.2	29.9	26.2	17.3	10.9	13.8	20.3
	MRT [59]	Def-DETR-R50	18.4	25.2	29.6	18.3	17.3	7.0	14.6	18.6
	2PC [25]	FRCNN-R50	37.4	31.3	42.5	17.0	27.9	28.6	22.1	29.5
	AT [32]	FRCNN-R50	32.1	26.4	41.7	35.5	27.6	27.4	19.7	30.1
	ALDI [24]	FRCNN-R50	39.8	36.0	54.1	44.0	35.5	27.9	26.6	37.7
SFDA	PETS [34]	FRCNN-VGG16	15.8	16.6	24.6	15.4	16.9	10.3	12.1	16.0
	LPLD [55]	FRCNN-R50	25.4	25.3	34.9	30.7	22.7	13.9	19.3	24.6
	MT [45]	FRCNN-R50	26.1	26.7	37.8	26.9	24.1	12.9	19.0	24.8
	DRU [26]	Def-DETR-R50	32.1	30.6	42.4	32.3	27.1	16.5	20.7	28.8
	SF-UT [14]	FRCNN-R50	30.8	33.1	48.0	37.9	29.0	14.5	23.6	31.0
	IRG [49]	FRCNN-R50	35.0	33.0	33.8	34.4	30.4	**31.8**	27.6	32.3
	DDT [15]	Def-DETR-R50	33.5	30.2	40.8	33.9	25.6	19.3	24.1	29.6
	FALCON [48]	FRCNN-R50	36.7	35.8	46.9	39.3	33.1	23.3	29.1	34.9
	CGSA [8]	RT-DETR-R50	40.4	35.8	48.5	44.1	32.9	30.0	25.5	36.7
	REUNION (ours)	FRCNN-R50	43.7	38.2	**53.9**	**45.7**	36.1	26.6	30.8	39.3
	REUNION (ours)	Def-DETR-R50	**45.5**	**39.6**	52.8	45.2	**36.3**	28.4	**32.5**	**40.0**

Table A4: Effect of Confidence Neighbors K.

K	AP^S	AP^M	AP^L
6	7.9	10.4	33.1
8	8.2	10.8	34.0
10	**8.4**	**11.1**	**34.6**
12	8.1	10.8	33.7
14	7.8	10.1	32.7

Table A5: Effect of Prototype Updating γ.

γ	AP^S	AP^M	AP^L
0.85	8.0	10.7	34.0
0.90	**8.4**	11.1	**34.6**
0.95	8.0	**11.4**	34.2
0.99	8.1	10.9	34.1

Increasing K to 10 yields the best performance, as a larger set of reliable neighbors forms a more stable semantic neighborhood that helps recover fragmented features caused by adverse weather and challenging light conditions. However, further increasing K (e.g., $K = 12$ or $K = 14$) slightly degrades performance, since overly large neighborhoods may introduce semantically mismatched connections and lead to negative transfer. Therefore, $K = 10$ provides the best trade-off between semantic richness and noise robustness.

A.4.2 Effect of Prototype Updating γ.

We maintain class- and weather-conditioned prototypes to provide global semantic anchors and mitigate pseudo-label drift. Table A5 reports the impact of the prototype updating momentum parameter γ, which balances the contribution of the current mini-batch against the historical moving average, as described in Section 3.5 of the main paper.

The results show that the adaptation process is sensitive to the choice of γ. A high value (e.g., $\gamma = 0.99$) causes the prototypes to update too aggressively from the current mini-batch. Under severe visibility degradation, this leads to unstable training because noisy pseudo-labels and weather-induced artifacts can drift the semantic anchors. In contrast, an excessively low value (e.g., $\gamma = 0.85$) makes the prototypes overly rigid, preserving historical representations and limiting the model's ability to adapt to the target weather and light conditions.

The best performance is achieved at intermediate values (e.g., $\gamma = 0.90$ or 0.95), where the prototypes stabilize training while remaining flexible enough to capture evolving weather- and light-invariant semantics. This balanced update particularly benefits large (AP^L) and medium (AP^M) objects, where consistent global semantic clustering is important for reliable domain adaptation.

Table A6: Robustness of REUNION to inaccurate weather labels on Cityscapes-Adverse. Weather labels are randomly corrupted on a fixed 50% subset of the unlabeled target adaptation data.

Weather-label noise	0%	10%	30%	50%
mAP	35.2	34.9	34.0	31.7
ΔmAP	0.0	-0.3	-1.2	-3.5

A.4.3 Robustness to Inaccurate Weather Labels.

To evaluate the robustness of REUNION to ambiguous or inaccurate weather annotations, we conduct a controlled label-corruption experiment on a fixed 50% subset of the unlabeled target adaptation data. We randomly corrupt 10%, 30%, and 50% of the weather labels by replacing each selected label with a uniformly sampled incorrect label from the remaining six weather conditions. The corrupted assignments are fixed throughout adaptation, while all other experimental settings and evaluation splits remain unchanged.

As shown in Table A6, REUNION achieves 35.2 mAP without label corruption and maintains 34.9, 34.0, and 31.7 mAP under 10%, 30%, and 50% label corruption, respectively. These results indicate gradual performance degradation, with only a 0.3-AP decrease at 10% noise and a 1.2-AP decrease at 30% noise.

Although incorrect weather labels can affect weather-aware hyperedges and class-weather prototypes, the complementary feature-based and intra-image hyperedges provide additional relational information that does not rely on weather labels. These findings suggest that REUNION is not entirely dependent on accurate weather-label assignments, although its performance gradually degrades with increasing label corruption.

A.5 Performance under Real-World Adverse Weather Conditions

To further evaluate the generalizability of REUNION beyond synthetically generated weather conditions, we conduct additional experiments on the real-world ACDC [43] dataset, which includes four adverse conditions: fog, rain, nighttime, and snow.

We adopt a source-free domain adaptation setting from Cityscapes [7] to ACDC [43], using a fixed 50% subset of the unlabeled target adaptation data. As shown in Table A7, REUNION achieves 36.2 mAP, outperforming the source-only baseline (17.8 mAP) and IRG [49] (25.7 mAP) by 18.4 and 10.5 AP points, respectively. Meanwhile, REUNION outperforms its competitor, IRG, across all four real adverse-weather conditions, with gains of 18.0, 8.0, 7.7, and 8.3 AP under fog, rain, nighttime, and snow, respectively. These results indicate that the effectiveness of REUNION extends beyond Cityscapes-derived synthetic weather variants. In addition, the improvement at night indicates that severe real-world illumination shifts and reflections remain challenging.

A.6 Performance Comparison to the state of the art VLMs and MFMs

Table A8 compares `REUNION` with existing SFDA vision specialists as well as recent Vision-Language Models (VLMs) and Multimodal Foundation Models (MFMs). Among the vision specialists, `REUNION` achieves the best performance on both small and large objects, reaching 8.4 AP^S and 34.6 AP^L. The improvement on large objects is particularly significant, indicating that the proposed hypergraph formulation effectively propagates semantic information across related proposals and stabilizes pseudo-label learning under severe visibility degradation. In contrast, general-purpose VLMs and MFMs exhibit substantially lower detection performance in this setting. While some models (e.g., Gemini 2.5 Flash and Qwen 3.0-VL) achieve good performance on large objects, their results on small and medium objects remain very limited. This gap highlights the difficulty of applying general multimodal models directly to dense detection tasks under adverse weather without task-specific adaptation. Overall, these results demonstrate that specialized SFDA methods remain more effective for domain-adaptive object detection, while `REUNION` further advances the state of the art by leveraging hypergraph-guided relational learning.

Fig. A2 presents additional qualitative comparisons between `REUNION` and several VLM/MFM baselines under diverse adverse-weather and light conditions. Consistent with the quantitative

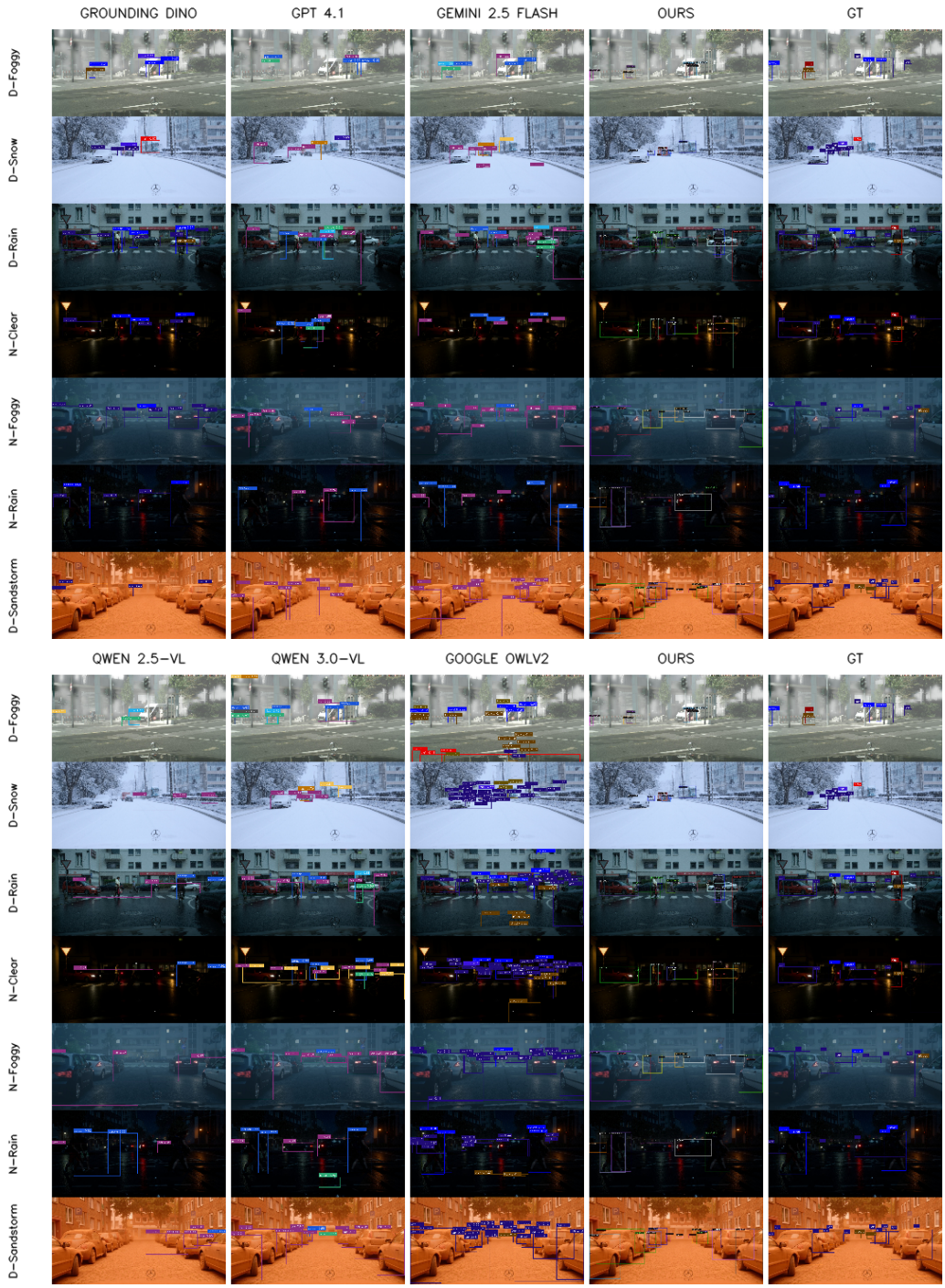


Figure A2: Qualitative results for the Cityscapes-Adverse dataset are displayed from left to right, showcasing the results from Grounding Dino, GPT-4.1, Gemini 2.5 Flash, Qwen 2.5-VL, Qwen 3.0-VL, Google OWLv2 and the proposed (Ours), with the ground truth (GT) presented on the far right.

Table A7: Performance comparison of source-free domain adaptive object detection from Cityscapes to the real-world ACDC dataset under four adverse conditions. Adaptation uses a fixed 50% subset of the unlabeled target data.

Method	Fog	Rain	Night	Snow	mAP
Source-Only	22.4	21.9	10.0	17.0	17.8
IRG [49]	31.3	28.7	20.8	22.0	25.7
REUNION (ours)	**49.3**	**36.7**	**28.5**	**30.3**	**36.2**

Table A8: Quantitative Comparison of Object Detection Capabilities. The performance of our proposed method is compared against state-of-the-art Source-Free Domain Adaptation (SFDA) vision specialists and modern Vision-Language Models (VLMs) / Multi-modal Foundation Models (MFMs).

Baselines	**Model**	AP^S	AP^M	AP^L
Vision Specialists	LPLD [55]	5.2	14.3	19.4
	MT [45]	5.3	**15.6**	19.8
	IRG [49]	7.4	12.4	29.4
	SF-UT [14]	6.7	11.8	28.0
	REUNION (**ours**)	**8.4**	11.1	**34.6**
VLMs / MFMs	Grounding Dino [35]	0.0	2.5	19.2
	GPT-4o [39]	0.0	0.0	0.0
	GPT-4.1 [40]	0.0	0.0	0.5
	Gemini 2.5 Flash [46]	0.1	4.9	29.9
	Qwen 2.5-VL [1]	0.0	0.0	0.0
	Qwen 3.0-VL [47]	0.1	5.4	31.6
	Google OWLv2 [38]	0.0	3.5	21.9

results in Table A8, REUNION produces more accurate and stable detections in challenging scenarios such as heavy rain, snow, fog, and nighttime environments. In particular, our method is able to correctly localize objects that are partially occluded or severely degraded by weather artifacts, whereas competing approaches often miss these instances or generate false positives in low-contrast regions.

Among the VLMs/MFMs, Qwen 2.5-VL largely fails to detect objects in the adverse-weather dataset. The newer Qwen 3.0-VL shows significant improvement and achieves performance close to REUNION. Similarly, Gemini 2.5 Flash demonstrates relatively strong detection capability and produces results comparable to our method in several cases. In contrast, GPT-4.1 fails to produce meaningful detection results. Although GPT-4.1 can often infer the presence of objects in the scene, its bounding-box localization is extremely inaccurate, resulting in nearly zero overall AP. This limitation has also been reported in prior studies [44, 6].

The visualization also reveals remaining failure cases in extremely challenging scenarios, such as nighttime rain, where strong reflections and severe illumination shifts continue to degrade detection reliability. Nevertheless, the overall qualitative results further demonstrate that REUNION improves the robustness and consistency of object detection across diverse adverse-weather conditions.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The main claims made in the abstract and introduction accurately reflect the paper's contributions and scope.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We discuss the limitations of the work in Section 5

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [N/A]

Justification: Our submission mainly focuses on application/architecture contributions and does not present formal theoretical results or proofs.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: We disclose all the information needed to reproduce the main experimental results of the paper.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in the supplemental material?

Answer: [Yes]

Justification: We provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results via an anonymous github link.

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

Answer:[Yes]

Justification: We provide the experimental settings, including data splits, training hyperparameters, model configurations, and evaluation protocols, in Sections 4 and A.1–A.4.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [No]

Justification: Following previous work, we mainly release results on the evaluation set.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).

- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We provide sufficient information on the computer resources for our experiments.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: Yes, our submission follows NeurIPS Code of Ethics.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: Yes, we discuss both potential positive societal impacts and negative societal impacts of the work in section Introduction and Future Work.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.

- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: Our submission poses no such risks.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: Yes, we properly credited and the license and terms of use explicitly mentioned and are properly respected.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.

- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

    Justification: Our submission does not release new assets.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.

- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: The core method development in this research does not involve LLMs as any important, original, or non-standard components.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.