BiDO: Bilevel Discriminative Learning with Online Synthesis for Object Detection in Foggy Scenes

Anonymous authors
Paper under double-blind review

Abstract

Object detection in foggy scenes is hampered by severe visibility degradation and the limited availability of trustworthy labels in real fog. Consequently, most methods rely on synthetic-to-real transfer, whose success depends on both the realism of synthesized fog and the detector's ability to generalize under inevitable distribution mismatch. We present BiDO, a bilevel discriminative learning framework that performs online synthesis while learning a joint dehazing-and-detection model. BiDO couples synthesis and task learning through bilevel optimization, using task-driven signals to adversarially regulate the synthesis process during training. By introducing stronger uncertainty via online synthesis, BiDO favors flatter solutions and yields more robust generalization to real fog. We evaluate BiDO on representative foggy-scene detection benchmarks under standard synthetic-to-real protocols and observe consistent improvements over prior state-of-the-art approaches. Additional algorithmic studies further highlight the advantages of online synthesis and demonstrate that BiDO can be integrated with different detector architectures. Code will be released if this work can be accepted.

1 Introduction

Vision-based object detection is a fundamental capability for safety-critical applications such as autonomous driving, intelligent transportation, and robotics. However, real-world deployment often involves adverse weather conditions, among which fog is particularly challenging due to atmospheric scattering and contrast attenuation. These effects blur object boundaries, suppress discriminative visual cues, and cause substantial performance degradation for detectors trained under clear-weather conditions. Meanwhile, high-quality annotations in foggy scenes are expensive and difficult to obtain, resulting in limited labeled data for training. Actually, the challenge of object detection in foggy scenes mainly stems from two fundamental challenges:

- ***Discriminative Degradation*.** Atmospheric scattering and contrast attenuation in foggy scenes substantially corrupt visual evidence, causing blurred object structures and weakened semantic cues. As a result, *the discriminative representations required for accurate object localization and recognition become unreliable, leading to degraded detection performance under adverse weather conditions.*
- ***Distribution Shift*.** The success of modern object detectors largely depends on large-scale annotated data, yet object detection in real foggy scenes are difficult to collect and annotate at scale. *A practical compromise is therefore to resort to synthetic fog or synthetic degradations for training, which inevitably introduces distribution shifts between synthetic training data and real-world foggy deployments, limiting generalization.*

Existing methods mainly address these challenges from two directions. Restoration-based pipelines aim to mitigate discriminative degradation by recovering image visibility before or during detection (Shao et al., 2020; Chen et al., 2021; Yang et al., 2022; Wu et al., 2023; Feng et al., 2024; Ma et al., 2025; Fu et al., 2025; Tsai et al., 2025). Transfer-based methods, such as domain adaptation and domain generalization, aim to reduce distribution shift by improving detector robustness under clear-to-fog or source-to-target transfer (Xu et al., 2023; Cao et al., 2023; Zhang et al., 2024; Kennerley et al., 2024; He et al., 2025a;b;c).

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

Despite their effectiveness, these two directions largely address the challenges in isolation. Restoration-based methods alleviate discriminative degradation by improving visibility, but their effectiveness is limited by how well restored images preserve task-relevant semantics. Transfer-based methods improve cross-domain robustness, but their effectiveness is limited by how well source or synthetic domains cover real-world fog variations. However, in real foggy scenes, the two challenges are inherently intertwined rather than independent, where discriminative degradation governs whether meaningful object semantics can be preserved under such conditions, while distribution shift determines whether the training data can faithfully reflect real-world fog characteristics.

Motivated by the above analysis, we propose BiDO (**Bi**level **D**iscriminative Learning with **O**nline Synthesis), a unified framework that integrates physics-constrained fog synthesis, image restoration, and object detection through bilevel optimization. By coupling online data generation with joint-task learning, BiDO enables the synthesis process to adapt to the evolving dehazing-and-detection model while generating physically constrained fog degradation. Specifically,

- To enable ***heterogeneous fog degradation***, we develop a global-to-local fog synthesis network that combines physics-based global fog generation with object-guided local density modulation. By incorporating depth information, spatial object priors, and physically constrained density increments, the proposed network enables adaptive control of both scene-level fog conditions and object-region degradation.
- To address ***distribution shift***, we develop a physics-inspired online fog synthesis mechanism that continuously adapts the generated fog distribution according to task feedback, improving its alignment with real-world fog variations.
- To achieve ***joint optimization***, we formulate online fog synthesis and joint dehazing-and-detection learning as a bilevel optimization problem. This formulation unifies physics-constrained fog generation, image restoration, and object detection, enabling the synthesis process to adapt to the evolving joint-task model by restoration and detection feedback.

2 Related Works

2.1 Restoration-based Methods

Fog-induced discriminative degradation originates from corrupted image formation, where scattering and attenuation suppress textures, boundaries, and object details. A natural solution is therefore to restore image visibility before or during detection. Dehazing-then-detection methods typically train restoration models on synthetic hazy data and transfer them to real foggy scenes, with recent advances improving real-world robustness through domain adaptation, self-supervised learning, generative priors, degradation-aware modeling, and target-domain haze synthesis (Shao et al., 2020; Chen et al., 2021; Yang et al., 2022; Wu et al., 2023; Feng et al., 2024; Ma et al., 2025; Fu et al., 2025; Tsai et al., 2025). Beyond this sequential paradigm, several methods integrate restoration into the detection pipeline via differentiable image processing (Liu et al., 2022; Kalwar et al., 2023), detector-guided restoration (Sun et al., 2022), unsupervised image translation (Wu & Jin, 2024), joint restoration-detection frameworks (Wang et al., 2024), and degradation-aware learning (Liu et al., 2025). While these methods improve visibility and robustness under adverse weather, they mainly optimize restoration quality or detector architectures, leaving the interaction between fog synthesis and detector learning underexplored.

2.2 Transfer-based Methods

Foggy-scene detection also suffers from a clear distribution gap between training and deployment domains, since large-scale annotated real foggy data are difficult to obtain. This motivates transfer-based methods that adapt detection knowledge from clear-weather or synthetic source domains to foggy target scenes. Existing methods improve cross-domain robustness through multi-view representation learning (Xu et al., 2023), object-level contrastive alignment (Cao et al., 2023), multi-granularity invariant learning (Zhang et al., 2024), category-aware knowledge transfer (Kennerley et al., 2024), and diffusion-based transferable priors (He et al., 2025a;b). More recently, discrepancy-aware adaptation further emphasizes hard-instance alignment under severe domain shifts (He et al., 2025c). Despite their effectiveness, these methods mainly reduce source-target discrepancies at the

model level and often rely on target-domain distributions or multiple source domains, which may limit robustness under unseen fog conditions.

3 BIDO

3.1 JOINT DEHAZING-AND-DETECTION

We formulate foggy-scene object detection as $\mathbf{z} = \mathcal{H}(\boldsymbol{\nu}; \mathbf{y})$, where $\mathbf{y}$ and $\mathbf{z}$ denote the foggy observation and detection output, respectively, and $\boldsymbol{\nu}$ parameterizes the detector. Given a clear image $\mathbf{x}$, a synthesis network $\mathcal{F}$ generates its foggy counterpart $\mathbf{y} = \mathcal{F}(\boldsymbol{\pi}; \mathbf{x})$. To mitigate fog-induced visual degradation and facilitate object detection, we introduce a dehazing network g and establish the following data flow: $\mathbf{x} \xrightarrow{\mathcal{F}(\boldsymbol{\pi};\cdot)} \mathbf{y} \xrightarrow{g(\boldsymbol{\mu};\cdot)} \hat{\mathbf{x}} \xrightarrow{\mathcal{H}(\boldsymbol{\nu};\cdot)} \mathbf{z}$, where $\hat{\mathbf{x}} = g(\boldsymbol{\mu}; \mathbf{y})$ denotes the restored image, and $\boldsymbol{\mu}$ parameterizes the dehazing network. Unlike conventional dehazing-then-detection pipelines that optimize the two tasks separately, we jointly optimize restoration and detection to learn task-relevant restored representations. For notational convenience, we define the joint-task parameters as $\boldsymbol{\theta} \triangleq [\boldsymbol{\mu}; \boldsymbol{\nu}]$.

3.2 CLEAR-TO-FOG SYNTHESIS NETWORK

We define a learnable clear-to-fog synthesis network $\mathcal{F}$ that generates paired training samples:

$$\mathbb{D}_{\boldsymbol{\pi}} \triangleq \{(\mathbf{x}_i, \mathbf{y}_i) \,|\, \mathbf{y}_i = \mathcal{F}(\boldsymbol{\pi}; \mathbf{x}_i)\}_{i=1}^{S}, \tag{1}$$

where $\mathbf{x}_i$ and $\mathbf{y}_i$ denote the clear image and its synthesized foggy counterpart, respectively. The synthesis network consists of global fog generation and object-guided local density modulation, parameterized by $\boldsymbol{\pi} = (\boldsymbol{\pi}_g, \boldsymbol{\pi}_l)$. Given a clear image and its depth information, the global module predicts scattering, atmospheric light, and imaging parameters to generate a base density field ρ_g and transmission map T_g, controlling scene-level fog degradation.

Based on the global fog field, the local module constructs a spatial prior from ground-truth bounding boxes and combines it with image and depth information to predict a density modulation map. A density constraint map, derived from the base fog conditions, depth information, and physical bounds, limits the allowable local density increment $\Delta\rho_l$ within and around object regions. The resulting density and transmission are given by

$$\rho(p) = \rho_g(p) + \Delta\rho_l(p), \; T(p) = T_g(p) \exp[-\beta D(p) \Delta\rho_l(p)], \tag{2}$$

where p denotes a pixel location, β is the scattering coefficient, and $D(p) = 2[1 - d(p)]$ is a relative distance factor derived from the normalized inverse depth $d(p)$. The local density increment introduces additional optical attenuation in object regions while retaining the global base fog field. The resulting transmission map is used to synthesize foggy images through an atmospheric scattering and imaging degradation process incorporating atmospheric light and optical blur (Wu et al., 2023).

Within the physically feasible synthesis space Π, we encourage pronounced fog degradation by maximizing the discrepancy between synthesized foggy images and their clear counterparts:

$$\max_{\boldsymbol{\pi} \in \Pi} \Phi(\boldsymbol{\pi}; \mathbb{C}), \tag{3}$$

where $\mathbb{C} \triangleq \{\mathbf{x}_i\}_{i=1}^{S}$ denotes the set of clear inputs. The discrepancy is defined as

$$\Phi(\boldsymbol{\pi}; \mathbb{C}) = \sum_{i=1}^{S} \|\mathcal{F}(\boldsymbol{\pi}; \mathbf{x}_i) - \mathbf{x}_i\|_1 . \tag{4}$$

This objective encourages stronger degradation subject to the constraints in Π. In the subsequent bilevel formulation, joint restoration and detection feedback further guides the synthesis process toward fog conditions that challenge the current task model.

3.3 TASK-GUIDED BILEVEL OPTIMIZATION FOR ONLINE FOG SYNTHESIS

Given the synthesized dataset $\mathbb{D}_{\boldsymbol{\pi}}$, a straightforward strategy is to optimize the joint dehazing-and-detection model under a fixed synthesis process:

$$\min_{\boldsymbol{\theta} \in \Theta} \Psi\big(\boldsymbol{\theta}; \mathbb{D}_{\boldsymbol{\pi}}\big), \tag{5}$$

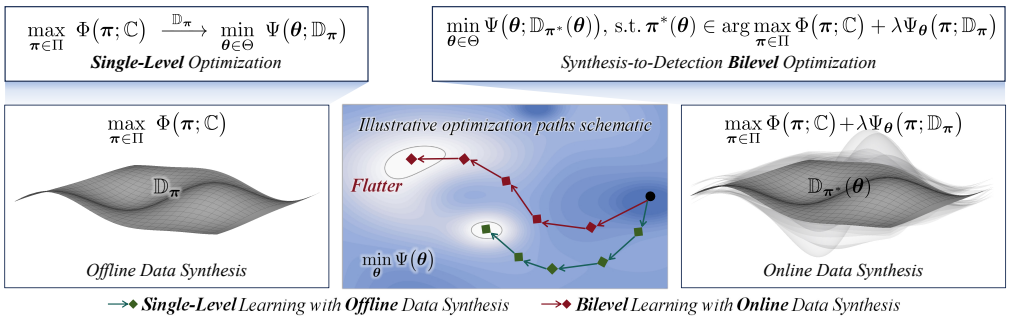


Figure 1: **Single-Level *vs.* Bilevel.** The single-level formulation first solves a synthesis objective to obtain a fixed synthetic dataset $\mathbb{D}_{\boldsymbol{\pi}}$, and then optimizes the joint-task parameters on this offline data (*Left*), which corresponds to the green trajectory in the optimization schematic and converges to a feasible solution. In contrast, the synthesis-to-detection bilevel formulation couples synthesis with task feedback and performs online synthesis during training (*Right*), yielding the red trajectory. By injecting larger uncertainty through online synthesis, bilevel learning encourages a flatter optimum and thus improves generalization.

where $\boldsymbol{\theta} = [\,\boldsymbol{\mu}; \boldsymbol{\nu}\,]$ denotes the joint-task parameters, and Ψ is the supervised objective for dehazing and detection. Specifically, we define$\Psi\big(\boldsymbol{\theta}; \mathbb{D}_{\boldsymbol{\pi}}\big) = \mathcal{L}_{\text{deh}}\big(\boldsymbol{\mu}; \mathbb{D}_{\boldsymbol{\pi}}\big) + \alpha\, \mathcal{L}_{\text{det}}\big(\boldsymbol{\nu}; \mathbb{D}_{\boldsymbol{\pi}}\big)$, where $\mathcal{L}_{\text{deh}}$ and $\mathcal{L}_{\text{det}}$ denote the dehazing and detection losses, respectively, and $\alpha > 0$ balances the two tasks.

Although Eq. 5 enables joint optimization of dehazing and detection, such an offline synthesis paradigm fixes the data generation process before model training, and therefore fails to capture the mutual dependency between fog synthesis and detector learning.

To model this interaction, BiDO formulates online synthesis and detector learning as a bilevel optimization problem. For a given synthesis process, the joint task model should minimize the dehazing and detection losses over the resulting foggy samples. Conversely, for a given task model, the synthesis network should adapt its degradation parameters to expose challenging fog conditions for the current model, while satisfying the physical constraints imposed by the clear-to-fog synthesis process. Based on this nested dependency, the resulting BiDO formulation is written as

$$\min_{\boldsymbol{\theta}\in\Theta} \Psi\big(\boldsymbol{\theta}; \mathbb{D}_{\boldsymbol{\pi}^*(\boldsymbol{\theta})}\big)\ ,\ \text{s.t.}\ \boldsymbol{\pi}^*(\boldsymbol{\theta}) \in \arg\max_{\boldsymbol{\pi}\in\Pi} \Phi(\boldsymbol{\pi}; \mathbb{C}) + \lambda \Psi_{\boldsymbol{\theta}}(\boldsymbol{\pi}; \mathbb{D}_{\boldsymbol{\pi}}), \tag{6}$$

where Π denotes the physically feasible synthesis space defined by the global and local fog-generation constraints, $\lambda > 0$ controls the contribution of task-level feedback, and $\Psi_{\boldsymbol{\theta}}$ denotes the joint-task loss evaluated with $\boldsymbol{\theta}$ fixed.

As shown in Fig. 1, this bilevel formulation couples task-adaptive online fog synthesis with joint-task learning under physical constraints, yielding an optimization trajectory distinct from conventional single-level learning with fixed offline synthesis. The optimization procedure is presented next.

3.4 Alternating Penalty-based Solver

Directly solving Eq. 6 is computationally expensive, as computing the exact hypergradient requires differentiating through the lower-level optimization process. To obtain a scalable first-order approximation, we adopt an alternating penalty-based strategy that couples online fog synthesis with joint-task learning without explicit second-order differentiation.

For notational convenience, we define

$$\begin{cases} h_{\boldsymbol{\theta}}(\boldsymbol{\theta}, \boldsymbol{\pi}) \triangleq \Psi\big(\boldsymbol{\theta}; \mathbb{D}_{\boldsymbol{\pi}}\big), \\ h_{\boldsymbol{\pi}}(\boldsymbol{\theta}, \boldsymbol{\pi}) \triangleq -\Phi\big(\boldsymbol{\pi}; \mathbb{C}\big) - \lambda \Psi\big(\boldsymbol{\theta}; \mathbb{D}_{\boldsymbol{\pi}}\big). \end{cases} \tag{7}$$

Here, $h_{\boldsymbol{\theta}}$ denotes the upper-level joint-task objective, while $h_{\boldsymbol{\pi}}$ negates the synthesis objective to convert the lower-level maximization into minimization. The bilevel problem can be written as

$$\min_{\boldsymbol{\theta}\in\Theta} h_{\boldsymbol{\theta}}\big(\boldsymbol{\theta}, \boldsymbol{\pi}^*(\boldsymbol{\theta})\big),\ \text{s.t.}\ \boldsymbol{\pi}^*(\boldsymbol{\theta}) \in \arg\min_{\boldsymbol{\pi}\in\Pi} h_{\boldsymbol{\pi}}(\boldsymbol{\theta}, \boldsymbol{\pi}). \tag{8}$$

Algorithm 1 Alternating Penalty-based Bilevel Optimization

Require: Initial variables $(\boldsymbol{\theta}_0, \boldsymbol{\pi}_0)$; outer iterations T; lower-level iterations M; step sizes $\eta_\pi, \alpha_\theta, \alpha_\pi$; penalty coefficient γ.
Ensure: Optimized joint-task and synthesis parameters $(\boldsymbol{\theta}_T, \boldsymbol{\pi}_T)$.

```
Initialize π̂_0^(0) = π_0.
for t = 0 to T − 1 do
    % Lower-level synthesis response approximation
    for m = 0 to M − 1 do
        Update π̂_t^(m+1) according to (11).
    end for
    Set π̂_t = π̂_t^(M).
    % First-order penalty-gradient estimation
    Compute g_t^θ and g_t^π according to (13).
    % Joint parameter update
    Update (θ_{t+1}, π_{t+1}) according to (14).
    % Warm start for the next lower-level approximation
    Set π̂_{t+1}^(0) = π̂_t.
end for
```

To avoid explicitly differentiating through the lower-level solution, we introduce a penalty objective based on the lower-level optimality gap:

$$\mathcal{P}_\gamma(\boldsymbol{\theta}, \boldsymbol{\pi}) = h_{\boldsymbol{\theta}}(\boldsymbol{\theta}, \boldsymbol{\pi}) + \gamma \left[h_{\boldsymbol{\pi}}(\boldsymbol{\theta}, \boldsymbol{\pi}) - v(\boldsymbol{\theta}) \right], \tag{9}$$

where $\gamma > 0$ is the penalty coefficient and

$$v(\boldsymbol{\theta}) \triangleq \min_{\boldsymbol{\pi}' \in \Pi} h_{\boldsymbol{\pi}}(\boldsymbol{\theta}, \boldsymbol{\pi}') \tag{10}$$

denotes the optimal value of the lower-level problem. The penalty term measures the deviation of the current synthesis objective from its lower-level optimum, encouraging the synthesis parameters to satisfy the lower-level optimality condition.

Since exactly evaluating $v(\boldsymbol{\theta})$ is computationally prohibitive, we approximate the lower-level solution using finitely many projected gradient steps. At iteration t, the task parameters $\boldsymbol{\theta}_t$ are fixed, and the auxiliary synthesis parameters are initialized as $\hat{\boldsymbol{\pi}}_t^{(0)}$. For $m = 0, \ldots, M-1$, we perform

$$\hat{\boldsymbol{\pi}}_t^{(m+1)} = \mathrm{Proj}_\Pi\Big(\hat{\boldsymbol{\pi}}_t^{(m)} - \eta_\pi \nabla_{\boldsymbol{\pi}} h_{\boldsymbol{\pi}}\big(\boldsymbol{\theta}_t, \hat{\boldsymbol{\pi}}_t^{(m)}\big) \Big), \tag{11}$$

where $\eta_\pi > 0$ is the lower-level step size and Proj_Π denotes projection onto the physically feasible synthesis space. The resulting approximation is denoted by

$$\hat{\boldsymbol{\pi}}_t \triangleq \hat{\boldsymbol{\pi}}_t^{(M)}. \tag{12}$$

Using $\hat{\boldsymbol{\pi}}_t$ to approximate the lower-level optimum, we construct the following first-order penalty-gradient estimators:

$$\begin{aligned} \boldsymbol{g}_t^\theta &= \nabla_{\boldsymbol{\theta}} h_{\boldsymbol{\theta}}\big(\boldsymbol{\theta}_t, \boldsymbol{\pi}_t\big) + \gamma \Big[\nabla_{\boldsymbol{\theta}} h_{\boldsymbol{\pi}}\big(\boldsymbol{\theta}_t, \boldsymbol{\pi}_t\big) - \nabla_{\boldsymbol{\theta}} h_{\boldsymbol{\pi}}\big(\boldsymbol{\theta}_t, \hat{\boldsymbol{\pi}}_t\big) \Big], \\ \boldsymbol{g}_t^\pi &= \nabla_{\boldsymbol{\pi}} h_{\boldsymbol{\theta}}\big(\boldsymbol{\theta}_t, \boldsymbol{\pi}_t\big) + \gamma \nabla_{\boldsymbol{\pi}} h_{\boldsymbol{\pi}}\big(\boldsymbol{\theta}_t, \boldsymbol{\pi}_t\big). \end{aligned} \tag{13}$$

Here, the gradients are evaluated by treating the current and auxiliary synthesis parameters as fixed when differentiating with respect to $\boldsymbol{\theta}$, thereby avoiding explicit backpropagation through the lower-level optimization trajectory.

Finally, the task and synthesis parameters are updated by

$$\boldsymbol{\theta}_{t+1} = \mathrm{Proj}_\Theta\big(\boldsymbol{\theta}_t - \alpha_\theta \boldsymbol{g}_t^\theta\big), \ \boldsymbol{\pi}_{t+1} = \mathrm{Proj}_\Pi\big(\boldsymbol{\pi}_t - \alpha_\pi \boldsymbol{g}_t^\pi\big), \tag{14}$$

where α_θ and α_π denote the corresponding learning rates.

Overall, the proposed solver alternates between approximating the task-dependent lower-level synthesis response and updating the joint-task and synthesis parameters by first-order penalty gradients. This procedure enables online generation of challenging foggy samples under physical constraints while coupling the synthesis process with dehazing-and-detection learning, without explicitly computing second-order hyper gradients. We summarize this procedure in Algorithm 1.

Table 1: Quantitative comparison with dehazing-then-detection pipelines on the RTTS dataset.

Task	Metrics	Input	DAD	PSD	D4	RIDCP	D4+	KANet	CoA	IPC	PHATNet	BiLaLoRA	BiDO
Dehazing	FADE↓	2.574	1.131	1.143	1.404	0.955	0.955	0.870	0.859	1.105	0.845	**0.752**	1.011
	PIQE↓	49.66	48.84	30.56	50.08	24.17	49.42	**20.68**	46.03	29.50	29.28	30.67	23.28
	BRISQUE↓	36.64	32.46	21.61	33.21	15.87	33.06	**11.88**	31.04	19.45	35.61	24.33	14.21
	MUSIQ↑	53.77	49.34	52.81	53.57	59.14	59.14	54.54	53.43	59.61	56.43	**61.77**	57.33
Detection	Person	0.464	0.452	0.465	0.468	0.462	0.468	0.456	**0.471**	0.468	0.446	0.468	0.468
	Car	0.362	0.379	0.372	0.373	0.378	0.373	0.357	0.387	**0.388**	0.371	0.385	0.378
	Bicycle	0.249	0.251	0.253	0.254	0.264	0.250	0.248	0.254	0.262	0.257	0.260	**0.316**
	Motorbike	0.201	0.205	0.220	0.218	0.247	0.224	0.219	0.234	0.233	0.217	0.240	**0.285**
	Bus	0.231	0.228	0.233	0.233	0.229	0.226	0.222	0.231	0.234	0.214	0.239	**0.240**
	mAP↑	0.301	0.303	0.309	0.309	0.316	0.308	0.300	0.315	0.317	0.301	0.318	**0.337**
	AP_{50} ↑	0.553	0.552	0.561	0.562	0.572	0.558	0.560	0.572	0.577	0.545	0.575	**0.601**
	AP_{75} ↑	0.292	0.291	0.299	0.301	0.307	0.303	0.286	0.304	0.309	0.293	0.312	**0.335**

4 Experiments

4.1 Experimental Settings

Training Settings. All experiments are implemented in PyTorch on four NVIDIA RTX 5090 GPUs. The joint dehazing-and-detection model is optimized using AdamW, with the learning rate decayed from 2×10^{-4} to 2×10^{-5}. The synthesis network is optimized using SGD with a learning rate of 2×10^{-5}. We set the task-feedback weight $\lambda = 0.55$ and penalty coefficient $\gamma = 2$. Input images are resized to a maximum side length of 640 pixels. Training augmentation includes random flipping and rotations (45° and 90°), while no test-time augmentation is applied.

Datasets. We evaluate BiDO on two real-world foggy benchmarks: RTTS Li et al. (2018) and Foggy Driving Sakaridis et al. (2018). For RTTS, we select 20K clear images from COCO as the source-domain training set, covering five categories: *Person*, *Car*, *Bicycle*, *Motorbike*, and *Bus*. Foggy training samples are generated online, while evaluation is conducted on real foggy RTTS images. For dehazing-then-detection comparisons, a Faster R-CNN with a ResNet-101 backbone is fine-tuned on the clear training images and then frozen during evaluation. The outputs of different dehazing methods are fed into this fixed detector under the same evaluation protocol. For Foggy Driving, we train on 2,975 synthetic foggy images from Foggy Cityscapes and evaluate on 101 real foggy images from Foggy Driving. Following Cityscapes, we consider eight categories: *Person*, *Rider*, *Car*, *Truck*, *Bus*, *Train*, *Motorcycle*, and *Bicycle*.

Compared Methods. We compare against recent SOTA end-to-end detectors for domain adaptation/generalization, including MIC (Hoyer et al., 2023), CMT (Cao et al., 2023), ALDI++ (Kay et al., 2025), DINO Teacher (Lavoie et al., 2025), IA-YOLO (Liu et al., 2022), DENet (Qin et al., 2023), FAT (Qin et al., 2025), and WARLearn (Agarwal et al., 2025) . We also benchmark representative dehazing-then-detection pipelines using a fixed Faster R-CNN (ResNet-101) detector, including DAD (Shao et al., 2020), PSD (Chen et al., 2021), D4 (Yang et al., 2022), RIDCP (Wu et al., 2023), D4+ (Yang et al., 2024), KANet (Feng et al., 2024), CoA (Ma et al., 2025), IPC (Fu et al., 2025), PHATNet (Tsai et al., 2025), and BiLaLoRA (Zhang et al., 2026).

Metrics. We evaluate detection using the COCO protocol (Lin et al., 2014) with mAP, computed by averaging AP over IoU thresholds from 0.50 to 0.95 in steps of 0.05. We also report AP_{50} and AP_{75}, which emphasize coarse detection and strict localization accuracy, respectively. For the auxiliary dehazing task, we adopt authoritative no-reference image quality metrics, including FADE (Choi et al., 2015), PIQE (Venkatanath et al., 2015), BRISQUE (Mittal et al., 2012), and MUSIQ (Ke et al., 2021), which measure haze density, perceptual quality, information content, and visual naturalness from complementary perspectives.

4.2 Compared with Dehazing-then-Detection Methods

Quantitative Comparisons. Table 1 compares BiDO with representative dehazing-then-detection pipelines on RTTS. In terms of restoration quality, BiDO remains competitive but does not dominate all no-reference metrics: it achieves relatively low PIQE and BRISQUE, while methods such as KANet and BiLaLoRA obtain better scores on FADE or MUSIQ. This difference is consistent with their stronger emphasis on standalone perceptual restoration, whereas BiDO treats dehazing as an

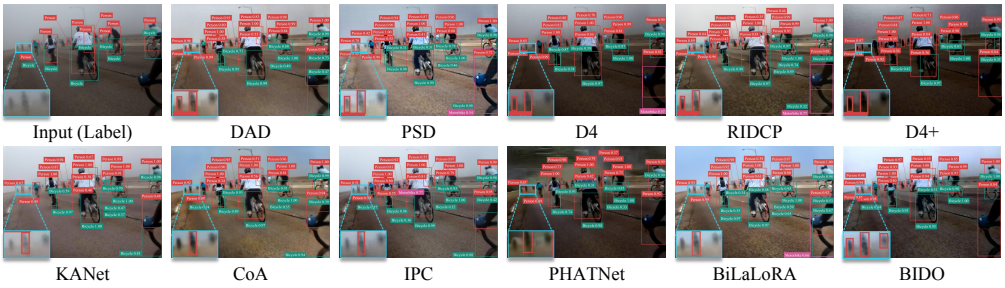


Figure 2: Qualitative comparisons on the RTTS dataset.

Table 2: Quantitative comparison of representative methods on the RTTS dataset.

Method	Person	Car	Bicycle	Motor.	Bus	mAP↑	AP_{50} ↑	AP_{75} ↑
DENet	0.373	0.308	0.243	0.184	0.117	0.245	0.444	0.239
IA-YOLO	0.340	0.252	0.202	0.124	0.058	0.195	0.383	0.178
FAT	0.287	0.304	0.216	0.061	0.077	0.189	0.333	0.183
WARLearn	0.376	0.314	0.242	0.167	0.133	0.246	0.444	0.239
BiDO	**0.468**	**0.378**	**0.316**	**0.285**	**0.240**	**0.337**	**0.601**	**0.335**

Table 3: Generalization results of representative methods on the Foggy Driving dataset.

Method	Bus	Bicycle	Car	Motor.	Person	Rider	Train	Truck	mAP↑	AP_{50} ↑	AP_{75} ↑
MIC	0.216	0.193	0.340	0.045	**0.156**	0.135	0.008	0.047	0.143	0.290	0.109
CMT	0.279	0.242	0.303	**0.119**	0.107	0.124	0.071	0.078	0.165	0.286	0.172
ALDI++	**0.398**	0.250	**0.352**	0.037	0.149	0.167	0.120	0.101	0.197	0.350	0.192
DINO Teacher	0.389	**0.310**	0.345	0.089	0.153	0.170	0.118	0.091	0.208	**0.360**	0.209
BiDO	**0.398**	0.252	0.345	0.089	**0.156**	**0.179**	**0.168**	**0.130**	**0.215**	0.355	**0.233**

intermediate task jointly optimized with downstream detection rather than explicitly maximizing image-quality scores.

This task-oriented optimization leads to a different performance profile on detection. BiDO achieves the best overall mAP, AP_{50}, and AP_{75}, with particularly large gains on Bicycle and Motorbike, while also improving Bus detection. These results indicate that better perceptual restoration does not necessarily translate into better recognition under fog. Instead, jointly optimizing restoration and detection encourages the restored representation to preserve information that is more useful for object discrimination and localization.

Qualitative Comparisons. Fig. 2 presents qualitative comparisons on the RTTS dataset. While several two-stage methods exhibit missed detections and inaccurate localization in fog-degraded regions, BiDO produces more complete detections with better-aligned bounding boxes, particularly for distant objects. These results suggest that joint dehazing-and-detection learning, coupled with task-guided online fog synthesis, helps preserve task-relevant visual information for robust detection under real-world fog.

4.3 Compared with End-to-End Detectors

Quantitative Comparisons. As shown in Table 2, BiDO achieves the highest detection performance across all evaluated metrics on the RTTS dataset. Specifically, BiDO achieves an mAP of 0.337, outperforming the strongest listed baseline, WARLearn (0.246), by 9.1 percentage points. BiDO also achieves AP_{50} and AP_{75} scores of 0.601 and 0.335, respectively, demonstrating improved detection accuracy under different IoU thresholds. Consistent improvements across all five object categories further indicate the effectiveness of BiDO in real-world foggy scenes.

Qualitative Comparisons. As illustrated in Fig. 3, BiDO achieves more accurate object localization and fewer missed detections than the compared methods in foggy scenes. These observations suggest that joint dehazing-and-detection learning, coupled with task-guided online fog synthesis, helps preserve discriminative visual information for object detection under fog degradation.

4.4 Generalization Verification

Table 3 shows that BiDO achieves the best overall detection performance on Foggy Driving, outperforming existing end-to-end detectors in mAP and AP_{75}. As illustrated in Fig. 4, BiDO produces more complete detections with better-aligned bounding boxes in fog-degraded regions, particularly for distant vehicles. These results demonstrate the effectiveness of BiDO in generalizing to unseen real-world foggy scenes.

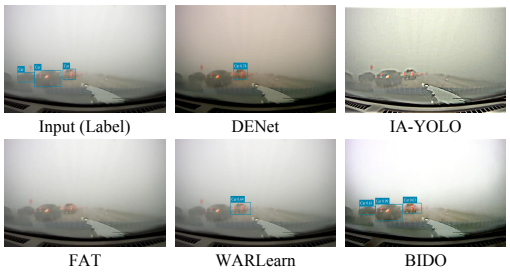


Figure 3: Qualitative comparisons of representative methods on the RTTS dataset.

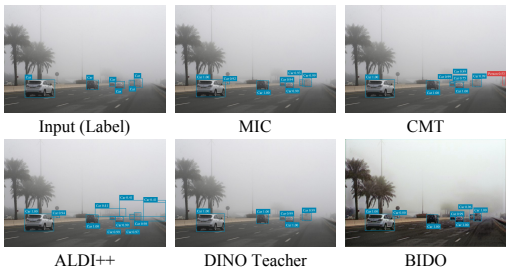


Figure 4: Qualitative comparisons of representative methods on the Foggy Driving dataset.

Table 4: Quantitative evaluation of architecture-agnostic flexibility on the RTTS dataset.

Dehazer	Model	Faster R-CNN (Ren et al., 2016)			YOLOX (Ge et al., 2021)			RT-DETR (Zhao et al., 2024)		
		mAP↑	AP_{50} ↑	AP_{75} ↑	mAP↑	AP_{50} ↑	AP_{75} ↑	mAP↑	AP_{50} ↑	AP_{75} ↑
DEA	Base	0.318	0.581	0.311	0.391	0.640	0.400	0.365	0.602	0.369
	BiDO	**0.337**$_{+0.019}$	**0.601**$_{+0.020}$	**0.335**$_{+0.024}$	**0.404**$_{+0.013}$	**0.656**$_{+0.016}$	**0.418**$_{+0.018}$	**0.381**$_{+0.016}$	**0.625**$_{+0.023}$	**0.387**$_{+0.018}$
MSBDN	Base	0.289	0.532	0.284	0.386	0.633	0.393	0.360	0.596	0.364
	BiDO	**0.322**$_{+0.033}$	**0.584**$_{+0.052}$	**0.319**$_{+0.035}$	**0.398**$_{+0.012}$	**0.650**$_{+0.017}$	**0.406**$_{+0.013}$	**0.374**$_{+0.014}$	**0.615**$_{+0.019}$	**0.378**$_{+0.014}$
DeHamer	Base	0.302	0.548	0.297	0.390	0.639	0.398	0.366	0.605	0.371
	BiDO	**0.330**$_{+0.028}$	**0.591**$_{+0.043}$	**0.326**$_{+0.029}$	**0.400**$_{+0.010}$	**0.652**$_{+0.013}$	**0.410**$_{+0.012}$	**0.378**$_{+0.012}$	**0.620**$_{+0.015}$	**0.383**$_{+0.012}$

5 Algorithmic Analysis

5.1 Architecture-Agnostic Flexibility

Table 4 evaluates BiDO on RTTS across three dehazing networks (DEA, MSBDN, and DeHamer) and three detectors (Faster R-CNN, YOLOX, and RT-DETR). Compared with conventional joint training on 20K fixed synthetic samples (Base), BiDO consistently improves mAP, AP_{50}, and AP_{75} across all nine dehazing-and-detection configurations. The mAP gains range from 0.010 to 0.033, with improvements observed consistently for both CNN-based and transformer-based detectors. Similar gains across different dehazing backbones further indicate that the benefit is not tied to a specific restoration architecture. These results demonstrate the architecture-agnostic flexibility of BiDO and its effectiveness across diverse restoration-detection combinations.

5.2 Effect of Global-to-Local Fog Synthesis

Fig. 5 visualizes how the global and local fog components evolve from their initial states through the task-guided bilevel optimization introduced in Section 3.3. The global component adjusts scene-wide fog conditions, providing a base degradation field, while the object-guided local component further modulates fog intensity and spatial distribution within and around target regions. Notably, the local variations exhibit spatially differentiated patterns, with degradation adjustments concentrated in specific object neighborhoods rather than uniformly distributed across the entire scene or object regions. This indicates that the synthesis network can adaptively regulate both the overall fog conditions and object-level degradation according to task feedback, enabling more targeted and challenging training-data generation under physical constraints.



Figure 5: **Visualization of Global-to-Local Fog Synthesis.** Initial fog (*Left*) and optimized fog obtained through task-guided bilevel optimization (*Right*). Orange and blue indicate increases and decreases relative to the reference state.

The quantitative results in Table 5 further confirm the benefit of local modulation. Across both single-level and bilevel optimization, object-guided local fog synthesis consistently improves detection over global-only degradation, indicating that spatially adaptive object-region modulation complements scene-level fog generation and provides more effective training conditions.

Table 5: Joint-task evaluation of different strategies.

Optimization		Haze Injection		Object Detection						Image Dehazing			
Single-Level	Bilevel	Global	Local	Person	Car	Bicycle	Motorbike	Bus	mAP↑	FADE↓	PIQE↓	BRISQUE↓	MUSIQ↑
✓	×	✓	×	0.456	0.351	0.268	0.220	0.220	0.303	1.186	26.03	16.17	55.02
✓	×	✓	✓	0.454	0.372	0.291	0.244	0.229	0.318	1.139	24.97	16.05	56.08
×	✓	✓	×	0.457	0.369	0.307	0.267	0.235	0.327	1.072	**22.80**	15.61	55.71
×	✓	✓	✓	**0.468**	**0.378**	**0.316**	**0.285**	**0.240**	**0.337**	**1.011**	23.28	**14.21**	**57.33**

Table 6: Quantitative results under different synthetic training set sizes on RTTS. Metrics are reported on a 0–1 scale; Δ denotes the signed change relative to the 20K setting.

Learning Strategy	20K Training Samples			10K Training Samples			5K Training Samples			1K Training Samples		
	mAP↑	AP_{50}↑	AP_{75}↑	mAP↑	AP_{50}↑	AP_{75}↑	mAP↑	AP_{50}↑	AP_{75}↑	mAP↑	AP_{50}↑	AP_{75}↑
Single-Level	0.318	0.581	0.311	0.296	0.549	0.289	0.268	0.510	0.258	0.231	0.451	0.216
Δ (vs. 20K)	–	–	–	-0.022	-0.032	-0.022	-0.050	-0.071	-0.053	-0.087	-0.130	-0.095
BiDO (Bilevel)	**0.337**	**0.601**	**0.335**	**0.328**	**0.587**	**0.325**	**0.316**	**0.568**	**0.312**	**0.300**	**0.548**	**0.295**
Δ (vs. 20K)	–	–	–	-0.009	-0.014	-0.010	-0.021	-0.033	-0.023	-0.037	-0.053	-0.040

5.3 Joint-Task Performance Evaluation

Table 5 also evaluates the effects of bilevel optimization on joint dehazing-and-detection performance. Compared with single-level training, bilevel optimization consistently improves overall detection accuracy under both fog synthesis configurations, highlighting the benefit of adapting training-data generation to the evolving joint-task model. The full BiDO configuration achieves the highest overall detection accuracy, with improvements across all evaluated object categories over the single-level baseline. Regarding image restoration, BiDO achieves the lowest FADE and highest MUSIQ among the evaluated configurations, although it does not consistently lead on all no-reference quality metrics. These results suggest that task-guided bilevel optimization supports downstream object recognition while maintaining competitive restoration quality, highlighting the benefit of coupling online synthesis with joint dehazing-and-detection learning.

5.4 Comparison across Synthetic Sample Sizes

Table 6 evaluates the sample efficiency of BiDO as the synthetic training set is progressively reduced from 20K to 1K samples. Across this process, BiDO consistently degrades more gracefully than single-level learning and maintains a clear advantage under increasingly data-scarce settings. Even with only 1K samples, BiDO remains competitive with substantially larger single-level training sets. These results indicate that task-guided online synthesis improves the utilization of limited training data and reduces reliance on large-scale fixed synthetic datasets.

5.5 Limitations

BiDO prioritizes task-oriented restoration for robust detection, and therefore does not consistently achieve the best scores on all perceptual dehazing metrics. This reflects a trade-off between visual restoration quality and downstream discriminability. Future work may incorporate unsupervised restoration constraints or perceptual consistency losses to further improve dehazing quality without compromising detection performance.

6 Conclusion

In conclusion, we proposed BiDO, a bilevel framework that couples online synthesis with joint dehazing-and-detection learning via task-driven signals. By regulating synthesis during training, BiDO injects stronger uncertainty and promotes flatter, more robust solutions that transfer better to real fog. Experiments on different synthetic-to-real benchmarks show consistent gains over prior state of the art, and further studies confirm the benefit of online synthesis and compatibility across detector architectures.

AI USE STATEMENT

Generative AI tools were used to assist with language editing, translation. They were not used to develop the research methodology, design or conduct experiments, analyze data, interpret results, or formulate scientific claims. All AI-assisted text and formatting changes were reviewed and verified by the authors. The authors take full responsibility for the final content of this work.

REFERENCES

Shubham Agarwal, Raz Birman, and Ofer Hadar. Warlearn: Weather-adaptive representation learning. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 4978–4987, 2025.

Shengcao Cao, Dhiraj Joshi, Liang-Yan Gui, and Yu-Xiong Wang. Contrastive mean teacher for domain adaptive object detectors. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 23839–23848, 2023.

Zeyuan Chen, Yangchao Wang, Yang Yang, and Dong Liu. Psd: Principled synthetic-to-real dehazing guided by physical priors. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 7180–7189, 2021.

Lark Kwon Choi, Jaehee You, and Alan Conrad Bovik. Referenceless prediction of perceptual fog density and perceptual image defogging. *IEEE Transactions on Image Processing*, 24(11): 3888–3901, 2015.

Yuxin Feng, Long Ma, Xiaozhe Meng, Fan Zhou, Risheng Liu, and Zhuo Su. Advancing real-world image dehazing: Perspective, modules, and training. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(12):9303–9320, 2024.

Jiayi Fu, Siyu Liu, Zikun Liu, Chun-Le Guo, Hyunhee Park, Ruiqi Wu, Guoqing Wang, and Chongyi Li. Iterative predictor-critic code decoding for real-world image dehazing. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 12700–12709, 2025.

Zheng Ge, Songtao Liu, Feng Wang, Zeming Li, and Jian Sun. Yolox: Exceeding yolo series in 2021. *arXiv preprint arXiv:2107.08430*, 2021.

Boyong He, Yuxiang Ji, Zhuoyue Tan, and Liaoni Wu. Boosting domain generalized and adaptive detection with diffusion models: Fitness, generalization, and transferability. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 1912–1923, 2025a.

Boyong He, Yuxiang Ji, Qianwen Ye, Zhuoyue Tan, and Liaoni Wu. Generalized diffusion detector: Mining robust features from diffusion models for domain-generalized detection. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 9921–9932, 2025b.

Xinyu He, Xinhui Li, and Xiaojie Guo. Differential alignment for domain adaptive object detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 17150–17158, 2025c.

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*, pp. 11721–11732, 2023.

Sanket Kalwar, Dhruv Patel, Aakash Aanegola, Krishna Reddy Konda, Sourav Garg, and K Madhava Krishna. Gdip: Gated differentiable image processing for object detection in adverse conditions. In *IEEE International Conference on Robotics and Automation*, pp. 7083–7089. IEEE, 2023.

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.

Junjie Ke, Qifei Wang, Yilin Wang, Peyman Milanfar, and Feng Yang. Musiq: Multi-scale image quality transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 5148–5157, 2021.

Mikhail Kennerley, Jian-Gang Wang, Bharadwaj Veeravalli, and Robby T Tan. Cat: Exploiting inter-class dynamics for domain adaptive object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16541–16550, 2024.

Marc-Antoine Lavoie, Anas Mahmoud, and Steven L. Waslander. Large self-supervised models bridge the gap in domain adaptive object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4692–4702, 2025.

Boyi Li, Wenqi Ren, Dengpan Fu, Dacheng Tao, Dan Feng, Wenjun Zeng, and Zhangyang Wang. Benchmarking single-image dehazing and beyond. *IEEE transactions on image processing*, 28 (1):492–505, 2018.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pp. 740–755. Springer, 2014.

Weifeng Liu, Jian Pang, Bingfeng Zhang, Jin Wang, Baodi Liu, and Dapeng Tao. See degraded objects: A physics-guided approach for object detection in adverse environments. *IEEE Transactions on Image Processing*, 34:2198–2212, 2025.

Wenyu Liu, Gaofeng Ren, Runsheng Yu, Shi Guo, Jianke Zhu, and Lei Zhang. Image-adaptive yolo for object detection in adverse weather conditions. In *Proceedings of the AAAI conference on artificial intelligence*, volume 36, pp. 1792–1800, 2022.

Long Ma, Yuxin Feng, Yan Zhang, Jinyuan Liu, Weimin Wang, Guang-Yong Chen, Chengpei Xu, and Zhuo Su. Coa: Towards real image dehazing via compression-and-adaptation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 11197–11206, 2025.

Anish Mittal, Anush Krishna Moorthy, and Alan Conrad Bovik. No-reference image quality assessment in the spatial domain. *IEEE Transactions on image processing*, 21(12):4695–4708, 2012.

Hongda Qin, Xiao Lu, Lucai Wang, and Yaonan Wang. Foggy-aware teacher: An unsupervised domain adaptive learning framework for object detection in foggy scenes. *IEEE Robotics and Automation Letters*, 10(7):7508–7515, 2025.

Qingpao Qin, Kan Chang, Mengyuan Huang, and Guiqing Li. Denet: Detection-driven enhancement network for object detection under adverse weather conditions. In Lei Wang, Juergen Gall, Tat-Jun Chin, Imari Sato, and Rama Chellappa (eds.), *Proceedings of the Asian Conference on Computer Vision*, pp. 491–507, 2023.

Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *IEEE transactions on pattern analysis and machine intelligence*, 39(6):1137–1149, 2016.

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

Yuanjie Shao, Lerenhan Li, Wenqi Ren, Changxin Gao, and Nong Sang. Domain adaptation for image dehazing. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 2808–2817, 2020.

Shangquan Sun, Wenqi Ren, Tao Wang, and Xiaochun Cao. Rethinking image restoration for object detection. *Advances in Neural Information Processing Systems*, 35:4461–4474, 2022.

Fu-Jen Tsai, Yan-Tsung Peng, Yen-Yu Lin, and Chia-Wen Lin. Phatnet: A physics-guided haze transfer network for domain-adaptive real-world image dehazing. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 5591–5600, 2025.

Narasimhan Venkatanath, Debbabi Praneeth, S Channappayya Sumohana, S Medasani Swarup, et al. Blind image quality evaluation using perception based features. In *2015 twenty first national conference on communications (NCC)*, pp. 1–6. IEEE, 2015.

Xiaofeng Wang, Xiao Liu, Hong Yang, Zhengyong Wang, Xiaoyue Wen, Xiaohai He, Linbo Qing, and Honggang Chen. Degradation modeling for restoration-enhanced object detection in adverse weather scenes. *IEEE Transactions on Intelligent Vehicles*, 2024.

Jiawei Wu and Zhi Jin. Unsupervised variational translator for bridging image restoration and high-level vision tasks. In *European Conference on Computer Vision*, pp. 214–231. Springer, 2024.

Rui-Qi Wu, Zheng-Peng Duan, Chun-Le Guo, Zhi Chai, and Chongyi Li. Ridcp: Revitalizing real image dehazing via high-quality codebook priors. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 22282–22291, 2023.

Mingjun Xu, Lingyun Qin, Weijie Chen, Shiliang Pu, and Lei Zhang. Multi-view adversarial discriminator: Mine the non-causal factors for object detection in unseen domains. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 8103–8112, 2023.

Yang Yang, Chaoyue Wang, Risheng Liu, Lin Zhang, Xiaojie Guo, and Dacheng Tao. Self-augmented unpaired image dehazing via density and depth decomposition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 2037–2046, 2022.

Yang Yang, Chaoyue Wang, Xiaojie Guo, and Dacheng Tao. Robust unpaired image dehazing via density and depth decomposition. *International Journal of Computer Vision*, 132(5):1557–1577, 2024.

Libo Zhang, Wenzhang Zhou, Heng Fan, Tiejian Luo, and Haibin Ling. Robust domain adaptive object detection with unified multi-granularity alignment. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(12):9161–9178, 2024.

Yan Zhang, Long Ma, Yuxin Feng, Zhe Huang, Fan Zhou, and Zhuo Su. Bilevel layer-positioning lora for real image dehazing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 29969–29978, 2026.

Yian Zhao, Wenyu Lv, Shangliang Xu, Jinman Wei, Guanzhong Wang, Qingqing Dang, Yi Liu, and Jie Chen. Detrs beat yolos on real-time object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16965–16974, 2024.