BiCo: Bilevel Collaborative Learning for Low-Light Vision

Anonymous authors
Paper under double-blind review

Abstract

Low-light conditions simultaneously impair visual quality and semantic perception, posing a fundamental challenge to robust vision in real-world environments. Rather than treating enhancement and high-level perception as isolated tasks, we propose a bilevel collaborative framework to explicitly coordinate pixel-level and semantic-level objectives for low-light vision. We first revisit the pixel-semantic gap and establish learnable role-balanced evaluators for the two levels, where pretrained foundation-model priors provide complementary quality and semantic guidance. Building upon these evaluators, we formulate the enhancer as the upper-level variable and the pixel- and semantic-level evaluators as lower-level variables, enabling reciprocal feedback between enhancement and perception through bilevel optimization. We further develop a two-stage learning strategy: cross-task collaboration learns a transferable enhancer through joint pixel-semantic optimization, while task-specific adaptation specializes the semantic network to a target downstream task with the enhancer fixed. The resulting enhancer captures task-agnostic meta-representations that generalize across different semantic tasks, including those unseen during collaborative learning. Extensive experiments on low-light detection and segmentation demonstrate consistent improvements in both visual quality and semantic performance, together with strong cross-task adaptability and generalization. The code will be released upon acceptance.

1 Introduction

Low-light images often suffer from poor visibility, low contrast, and amplified noise due to insufficient illumination. These degradations not only impair visual quality, but also obscure semantic information required for downstream understanding. In practice, some applications require seeing clearly, *e.g.,* enhancing nighttime traffic footage to recover license plates and road signs for visual inspection, whereas others require understanding accurately, *e.g.,* detecting pedestrians and vehicles for autonomous driving. Such dual demands have motivated extensive research on low-light vision.

Existing studies often address low-light vision from only one side of the pixel-semantic relationship. Pixel-level methods (Guo et al., 2020; Shi et al., 2024; Brateanu et al., 2026) improve visual quality but may distort task-relevant structures, while semantic-level methods (Cui et al., 2021; Du et al., 2024; Shin et al., 2026) enhance downstream perception but underuse cues from restored representations. Recent joint methods (Xue et al., 2022; Hashmi et al., 2023; Zhao et al., 2025) attempt to unify enhancement and perception within a shared framework. However, they still lack an explicit formulation of how restoration and perception should mutually constrain and benefit each other.

To bridge the pixel-semantic gap in low-light vision, we revisit the interaction between pixel-level restoration and semantic-level perception, highlighting the need for explicit cross-task collaboration beyond one-way coupling. Accordingly, we propose BiCo, a bilevel collaborative framework for jointly modeling the two levels of vision. As illustrated in Fig. 1, BiCo follows a two-stage design. Stage I learns a visual- and semantic-aware enhancer through bilevel pixel-semantic collaboration, while Stage II transfers the learned enhancer to task-specific adaptation for both seen and unseen downstream tasks. Quantitative results show that BiCo consistently benefits semantic perception while preserving high-quality enhancement, enabling robust low-light vision across different tasks. Our main contributions can be summarized as follows:

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.

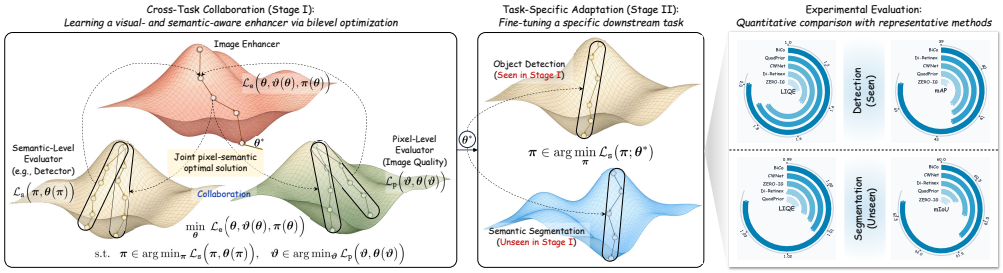

Figure 1: Overview of BiCo. The framework consists of Stage I (Left), which learns a visual- and semantic-aware enhancer via bilevel cross-task collaboration, and Stage II (Middle), which adapts the semantic network to a specific downstream task with the enhancer fixed. The Right panel presents quantitative comparisons with representative methods on different task settings.

- We revisit the pixel-semantic interaction in low-light vision from both forward and backward perspectives, revealing the imbalance induced by one-way coupling. Accordingly, we design role-balanced learnable evaluators for the two levels: a VLM-guided pixel-level evaluator and a VFM-guided semantic-level evaluator, providing complementary quality and semantic feedback for balanced cross-level collaboration.
- Building upon these evaluators, we propose BiCo, a bilevel collaborative framework that coordinates enhancement and semantic perception through bilevel optimization toward a joint pixel-semantic optimum. To solve this formulation, we derive a two-stage collaborative-adaptive strategy, which further enables the learned enhancer to transfer to both seen and unseen downstream tasks. *To the best of our knowledge, BiCo is the first work to explore pixel-semantic collaboration via bilevel optimization for low-light vision.*
- Extensive experiments on low-light enhancement, object detection, and semantic segmentation demonstrate consistent gains in both visual quality and semantic perception. Further analyses verify the necessity of the bilevel collaborative design and reveal strong cross-task adaptability and transferable meta-representation capability of the learned enhancer.

2 Related Work

Pixel-Level Vision in the Dark. Due to the difficulty of acquiring realistic paired low-light data, unsupervised and zero-reference enhancement has become a mainstream paradigm. Representative methods include EnGAN (Jiang et al., 2021), ZeroDCE (Guo et al., 2020), RUAS (Liu et al., 2021b), and later zero-reference frameworks (Liang et al., 2022; Ma et al., 2022; Shi et al., 2024). Recent studies further incorporate physical priors (Huang et al., 2026; Brateanu et al., 2026) and generative priors (Cho et al., 2026; Hua et al., 2026) to improve fidelity and robustness. However, these methods remain largely pixel-oriented and weakly coupled with downstream perception. Although some works introduce semantic guidance (Zheng et al., 2023; Wu et al., 2023; Liang et al., 2022), the interaction between enhancement and high-level vision remains insufficiently explored.

Semantic-Level Vision in the Dark. Existing semantic-level methods for vision in the dark mainly target object detection (Cui et al., 2021; Zhang et al., 2024; Du et al., 2024; Shin et al., 2026) and semantic segmentation (Wu et al., 2021; Lengyel et al., 2021; Lee et al., 2023). Early approaches rely on exposure modeling and image translation (Liang et al., 2021; Du et al., 2024), followed by domain adaptation and intermediate-domain learning (Wang et al., 2021; Dai & Van Gool, 2018; Sakaridis et al., 2019; 2020; Wu et al., 2021; Lengyel et al., 2021; Chen et al., 2026). Recent methods further explore self-guidance (Shin et al., 2026), frequency modeling (Sun et al., 2025a), and semantic enrichment (Li et al., 2026). However, they largely overlook the interaction between semantic perception and pixel-level enhancement.

Joint Pixel-Semantic Vision in the Dark. Recent studies have begun to jointly optimize pixel-level and semantic-level vision. Xue *et al.* (Xue et al., 2022) adopt alternating optimization in a unified multi-task framework, FeatEnHancer (Hashmi et al., 2023) injects enhancement-aware features into downstream perception, and Zhao *et al.* (Zhao et al., 2025) study restoration-detection synergy from a Lipschitz-continuity perspective. However, these methods still lack explicit modeling of cross-level interdependence, limiting effective pixel-semantic collaboration.

3 METHOD

3.1 RETHINKING LOW-LIGHT VISION

Low-Light Vision as Joint Pixel-Semantic Learning. Unlike conventional vision tasks that typically assume reliable image observations, low-light vision begins with intrinsically unreliable data. Therefore, pixel-level restoration is not merely a preprocessing choice, but an intrinsic requirement for recovering reliable visual evidence from degraded observations. Meanwhile, low-light vision cannot stop at visual restoration, since practical systems ultimately require semantic-level understanding for external tasks. The two levels are not independent. Better restoration provides clearer and more natural evidence for semantic analysis, while semantic perception identifies which structures and details are truly task-relevant. *This mutual dependence motivates low-light vision as a joint pixel-semantic learning problem, where pixel-level restoration and semantic-level perception are learned collaboratively rather than in isolation.* Based on this joint view, let $\mathcal{F}(\cdot;\boldsymbol{\theta})$ and $\mathcal{G}(\cdot;\boldsymbol{\pi})$ denote the pixel-level enhancer and semantic-level network parameterized by $\boldsymbol{\theta}$ and $\boldsymbol{\pi}$, respectively. Given a low-light image $\mathbf{y}$, the low-light vision pipeline is formulated as

$$\mathbf{x} = \mathbf{y} \oslash \mathcal{F}(\mathbf{y};\boldsymbol{\theta}), \quad \mathbf{s} = \mathcal{G}(\mathbf{x};\boldsymbol{\pi}), \tag{1}$$

where $\mathcal{F}(\mathbf{y};\boldsymbol{\theta})$ estimates the illumination under Retinex theory (Land & McCann, 1971), $\oslash$ denotes element-wise division, and $\mathbf{x}$ and $\mathbf{s}$ denote the enhanced image and semantic prediction, respectively.

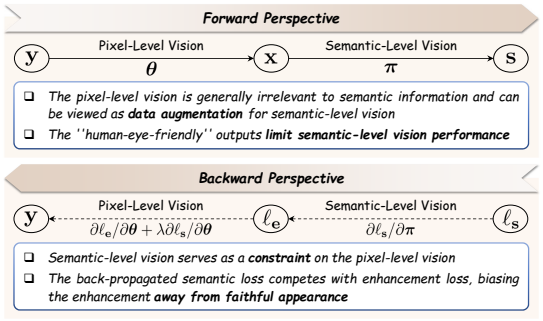

Figure 2: Interpretation of different perspectives.

Forward Perspective. From the forward propagation viewpoint, as illustrated in the top panel of Fig. 2, pixel-level vision is typically positioned as a preprocessing stage before semantic-level perception. In this cascaded pipeline, the enhancer $\boldsymbol{\theta}$ first transforms the degraded input $\mathbf{y}$ into an enhanced image $\mathbf{x}$, which is subsequently fed into the semantic model $\boldsymbol{\pi}$ for downstream prediction. Such a design essentially treats pixel-level enhancement as a form of data refinement or augmentation for semantic-level vision, without explicitly accounting for the semantic information required by the downstream task. As a result, enhancement is largely optimized toward human-perceived visual quality, while such "human-eye-friendly" outputs are not necessarily optimal for semantic perception and may even limit downstream performance.

Backward Perspective. From the optimization viewpoint, as illustrated in the bottom panel of Fig. 2, the semantic-level objective is back-propagated through the semantic network to provide task-driven supervision for the enhancer. While such feedback makes enhancement more responsive to downstream perception, it also introduces an asymmetric optimization process, where semantic-level vision acts as a constraint on pixel-level restoration. In particular, the back-propagated semantic loss may compete with the enhancement objective, leading to conflicting optimization directions for the enhancer. Consequently, the enhancer tends to favor task-specific semantic discrimination at the expense of faithful appearance restoration, thereby hindering a balanced collaboration between pixel-level and semantic-level vision.

Table 1: Evaluation of different perspectives.

Task	Metric	Perspective	Input	ZeroDCE	SCI	ZeroIG
Detection (TOOD)	CLIPIQA	Forward	0.449	0.388	0.432	0.271
		Backward		0.312	0.430	0.194
	mAP	Forward	42.98	42.29	42.15	39.66
		Backward		42.01	42.28	39.58
Segmentation (Segformer)	CLIPIQA	Forward	0.236	0.211	0.230	0.242
		Backward		0.275	0.218	0.175
	mIoU	Forward	62.91	62.38	60.91	60.47
		Backward		62.46	61.50	60.84

As shown in Table 1, improving visual quality does not necessarily translate into consistent gains in semantic performance. As shown in Table 1, improvements in semantic performance are often accompanied by degradation of visual quality, revealing an inherent conflict between the two objectives. Together, these two perspectives reveal an inherent asymmetry in the cascaded paradigm: pixel-level restoration is treated as a prerequisite in the forward pass, yet is dominated by semantic supervision in the backward optimization. *Such unequal roles prevent genuine pixel-semantic collaboration, motivating the need for balanced joint constraints.*

3.2 Role-Balanced Pixel-Semantic Evaluator Design

To address the unequal roles discussed in Sec. 3.1, we introduce learnable pixel-level and semantic-level evaluators with balanced optimization roles, establishing an equal-footing basis for subsequent pixel-semantic collaboration. Meanwhile, pretrained foundation-model priors are incorporated into the two evaluators to further strengthen their quality and semantic assessment capabilities.

Given the enhanced image $\mathbf{x}$, we first construct the basic representations for the two evaluators:

$$\mathbf{P} = \mathcal{H}(\mathbf{B}), \qquad \mathbf{S} = \mathcal{A}(\mathbf{F}), \tag{2}$$

where $\mathbf{P}$ and $\mathbf{S}$ denote the basic representations for the pixel-level and semantic-level evaluators, respectively. The features $\mathbf{B}$ and $\mathbf{F}$ are extracted from the backbone and the corresponding FPN of the task-specific semantic network, respectively. The mappings $\mathcal{H}$ and $\mathcal{A}$ denote the learnable pixel-level evaluator and lightweight projection module, respectively.

VLM-Guided Pixel-Level Evaluator. We summarize $\mathbf{P}$ as a compact quality descriptor $\mathbf{q} = [\mu(\mathbf{P}); \sigma(\mathbf{P})]$ and compare it with clear and low-light reference representations $\mathbf{v}^+$ and $\mathbf{v}^-$, yielding $d^{\pm} = \mathbf{q}^\top \mathbf{v}^{\pm}$. To complement this representation-level quality cue, we further leverage textual priors from a frozen CLIP model (Radford et al., 2021) for cross-modal quality supervision. Given textual prototypes $\mathbf{t}^+$ and $\mathbf{t}^-$ describing clear and low-light appearances, respectively, we define the quality direction as $\mathbf{r} = \mathbf{t}^+ - \mathbf{t}^-$. For the CLIP image representations $\mathbf{c_x}$, $\mathbf{c}^+$, and $\mathbf{c}^-$ of the enhanced image, clear unpaired reference, and low-light input, their projections along the quality direction are defined as $z_{\mathbf{x}} = \mathbf{c}_{\mathbf{x}}^\top \mathbf{r}$ and $z^{\pm} = (\mathbf{c}^{\pm})^\top \mathbf{r}$. We then measure appearance consistency by $i^{\pm} = 1 - \frac{1}{2}|z_{\mathbf{x}} - z^{\pm}|$, while the direct image-text consistency is given by $t_{\mathbf{x}}^{\pm} = \mathbf{c}_{\mathbf{x}}^\top \mathbf{t}^{\pm}$. The resulting objective is formulated as

$$\mathcal{L}_{\mathtt{p}} = \frac{1}{B} \sum_{b=1}^{B} \mathrm{softplus}\left(\frac{e_b^- - e_b^+ + m}{\tau_q} \right), \tag{3}$$

where $e^{\pm} = d^{\pm} + \alpha i^{\pm} + \beta t_{\mathbf{x}}^{\pm}$ denotes the fused quality score, with α and β balancing different quality cues. The parameters m and τ_q denote the margin and temperature, respectively.

VFM-Guided Semantic-Level Evaluator. We introduce a frozen DINO model (Siméoni et al., 2025) to extract the semantic representation $\mathbf{V} = \mathcal{V}_h(\mathbf{x})$ and establish semantic regularization between the task-specific representation $\mathbf{S}$ and the DINO-derived representation $\mathbf{V}$ through a bidirectional spatial contrastive objective:

$$\mathcal{L}_{\mathtt{r}} = -\frac{1}{2M} \sum_{i=1}^{M} \left[\log \frac{\exp\left(\mathrm{sim}(\mathbf{s}_i, \mathbf{v}_i)/\tau_s\right)}{\sum_{j=1}^{M} \exp\left(\mathrm{sim}(\mathbf{s}_i, \mathbf{v}_j)/\tau_s\right)} + \log \frac{\exp\left(\mathrm{sim}(\mathbf{v}_i, \mathbf{s}_i)/\tau_s\right)}{\sum_{j=1}^{M} \exp\left(\mathrm{sim}(\mathbf{v}_i, \mathbf{s}_j)/\tau_s\right)} \right], \tag{4}$$

where $\{\mathbf{s}_i\}_{i=1}^{M}$ and $\{\mathbf{v}_i\}_{i=1}^{M}$ denote the spatial features of $\mathbf{S}$ and $\mathbf{V}$, respectively, $\mathrm{sim}(\cdot,\cdot)$ denotes cosine similarity, and τ_s is the temperature. Corresponding locations form positive pairs, while the remaining locations serve as negatives, encouraging the task-specific representation to preserve transferable semantic structures. The final semantic-level evaluator combines task-specific loss ($\mathcal{L}_{\mathtt{t}}$) with VFM-guided regularization: $\mathcal{L}_{\mathtt{s}} = \mathcal{L}_{\mathtt{t}} + \lambda_{\mathtt{r}} \mathcal{L}_{\mathtt{r}}$, where $\lambda_{\mathtt{r}}$ balances the regularization term.

3.3 Bilevel Modeling with Pixel-Semantic Evaluators

Following Sec. 3.2, we obtain learnable pixel-level and semantic-level evaluators that provide complementary visual and semantic feedback for enhancement. The framework therefore involves three groups of learnable parameters: the enhancer parameters $\boldsymbol{\theta}$, pixel-level evaluator parameters $\boldsymbol{\vartheta}$, and semantic-level evaluator parameters $\boldsymbol{\pi}$. Since the optimal evaluators depend on the enhanced representations, while the enhancer should in turn be updated according to their feedback, these parameters exhibit an inherent hierarchical dependency. Accordingly, we treat the enhancer parameters as upper-level variables that shape the optimization of the two evaluators, leading naturally to a bilevel formulation (Liu et al., 2021a), where the evaluators define the lower-level objectives and the enhancer is optimized at the upper level through their collaborative feedback:

$$\min_{\boldsymbol{\theta}} \mathcal{L}_{\mathtt{e}}\Big(\boldsymbol{\theta}, \boldsymbol{\vartheta}(\boldsymbol{\theta}), \boldsymbol{\pi}(\boldsymbol{\theta})\Big), \ s.t., \begin{cases} \boldsymbol{\vartheta} \in \arg\min_{\boldsymbol{\vartheta}} \mathcal{L}_{\mathtt{p}}\Big(\boldsymbol{\vartheta}, \boldsymbol{\theta}(\boldsymbol{\vartheta})\Big), \\ \boldsymbol{\pi} \in \arg\min_{\boldsymbol{\pi}} \mathcal{L}_{\mathtt{s}}\Big(\boldsymbol{\pi}, \boldsymbol{\theta}(\boldsymbol{\pi})\Big), \end{cases} \tag{5}$$

where $\mathcal{L}_{\mathtt{e}}$, $\mathcal{L}_{\mathtt{p}}$, and $\mathcal{L}_{\mathtt{s}}$ denote the intrinsic enhancement loss, pixel-level evaluator, and semantic-level evaluator, respectively. In practice, $\mathcal{L}_{\mathtt{e}}$ adopts the unsupervised Retinex-based enhancement objective in (Liu et al., 2021b), while the task-specific component $\mathcal{L}_{\mathtt{t}}$ in $\mathcal{L}_{\mathtt{s}}$ is instantiated according to the downstream task, *e.g.,* cross-entropy for semantic segmentation (Long et al., 2015).

3.4 Bilevel Collaborative Learning

As defined in Sec. 3.2, the learnable pixel-level evaluator is constructed upon intermediate representations produced by the semantic-level network. Consequently, its optimization is partially coupled with the semantic-level evaluator through shared learnable parameters. We therefore decompose the semantic-level evaluator parameters as

$$\boldsymbol{\pi} = \boldsymbol{\kappa} \cup \boldsymbol{\vartheta}, \tag{6}$$

where $\boldsymbol{\vartheta}$ denotes the shared parameters involved in both pixel-level and semantic-level evaluation, while $\boldsymbol{\kappa}$ represents the remaining task-specific parameters of the semantic-level evaluator.

Algorithm 1 BiCo

Require: Initial parameters $\boldsymbol{\theta}_0, \boldsymbol{\pi}_0 = \boldsymbol{\kappa}_0 \cup \boldsymbol{\vartheta}_0$, step sizes $\eta_{\boldsymbol{\theta}}, \eta_{\boldsymbol{\vartheta}}, \eta_{\boldsymbol{\kappa}}, \eta_{\boldsymbol{\pi}}$, iterations T, K, and M.
Ensure: The enhancer $\boldsymbol{\theta}^*$ and semantic model $\boldsymbol{\pi}^*$.

```
% Cross-Task Collaboration
for t = 0, ..., T − 1 do
    for k = 0, ..., K − 1 do
        ϑ_t^{k+1} = ϑ_t^k − η_ϑ ∇_ϑ L_p(θ_t, ϑ_t^k, κ_t^k).
        κ_t^{k+1} = κ_t^k − η_κ ∇_κ L_s(θ_t, ϑ_t^{k+1}, κ_t^k).
    end for
    ϑ_{t+1} = ϑ_t^K,   κ_{t+1} = κ_t^K.
    Compute J(θ_t) by Equation 10.
    θ_{t+1} = Π_Θ (θ_t − η_θ ∇_θ J(θ_t)) .
end for
% Task-Specific Adaptation
Freeze θ* = θ_T and initialize π^0 = κ_T ∪ ϑ_T.
for m = 0, ..., M − 1 do
    π^{m+1} = π^m − η_π ∇_π ℓ_s(π^m, θ*).
end for
Return θ*, π* = π^M.
```

Based on the above parameter decomposition, we further develop a two-stage strategy to solve the bilevel formulation: Cross-Task Collaboration (CTC) for joint pixel-semantic optimization, followed by Task-Specific Adaptation (TSA) for downstream task adaptation.

Cross-Task Collaboration (CTC). Given the enhancer parameter $\boldsymbol{\theta}$, the lower-level optimization is solved by alternating between the pixel-level and semantic-level constraints:

$$\begin{cases} \boldsymbol{\vartheta}^{k+1} = \boldsymbol{\vartheta}^k - \eta_{\boldsymbol{\vartheta}} \nabla_{\boldsymbol{\vartheta}} \mathcal{L}_{\mathtt{p}}\big(\boldsymbol{\theta}, \boldsymbol{\vartheta}^k, \boldsymbol{\kappa}^k\big), \\ \boldsymbol{\kappa}^{k+1} = \boldsymbol{\kappa}^k - \eta_{\boldsymbol{\kappa}} \nabla_{\boldsymbol{\kappa}} \mathcal{L}_{\mathtt{s}}\big(\boldsymbol{\theta}, \boldsymbol{\vartheta}^{k+1}, \boldsymbol{\kappa}^k\big), \end{cases} \tag{7}$$

where $\eta_{\boldsymbol{\vartheta}}$ and $\eta_{\boldsymbol{\kappa}}$ are learning rates. This alternating process preserves the two lower-level constraints while enabling parameter sharing between pixel-level and semantic-level evaluators.

To optimize the upper-level enhancer without explicitly differentiating through the full alternating trajectory, we adopt the Value-function-based Projected Bilevel Gradient Descent (V-PBGD) (Shen & Chen, 2023) framework. We define two value functions for the lower-level constraints:

$$\phi_{\mathtt{p}}(\boldsymbol{\theta}, \boldsymbol{\kappa}) = \min_{\boldsymbol{\vartheta}} \mathcal{L}_{\mathtt{p}}\big(\boldsymbol{\theta}, \boldsymbol{\vartheta}, \boldsymbol{\kappa}\big), \quad \phi_{\mathtt{s}}(\boldsymbol{\theta}, \boldsymbol{\vartheta}) = \min_{\boldsymbol{\kappa}} \mathcal{L}_{\mathtt{s}}\big(\boldsymbol{\theta}, \boldsymbol{\vartheta}, \boldsymbol{\kappa}\big). \tag{8}$$

The corresponding value-function constraints are

$$\mathcal{L}_{\mathtt{p}}(\boldsymbol{\theta}, \boldsymbol{\vartheta}, \boldsymbol{\kappa}) - \phi_{\mathtt{p}}(\boldsymbol{\theta}, \boldsymbol{\kappa}) \le 0, \quad \mathcal{L}_{\mathtt{s}}(\boldsymbol{\theta}, \boldsymbol{\vartheta}, \boldsymbol{\kappa}) - \phi_{\mathtt{s}}(\boldsymbol{\theta}, \boldsymbol{\vartheta}) \le 0. \tag{9}$$

Accordingly, the upper-level optimization is relaxed as the following penalized objective:

$$\min_{\boldsymbol{\theta}} \quad \mathcal{J}(\boldsymbol{\theta}) = \mathcal{L}_{\mathtt{e}}\big(\boldsymbol{\theta}, \boldsymbol{\vartheta}, \boldsymbol{\kappa}\big) + \rho_{\mathtt{p}}\Big[\mathcal{L}_{\mathtt{p}}\big(\boldsymbol{\theta}, \boldsymbol{\vartheta}, \boldsymbol{\kappa}\big) - \phi_{\mathtt{p}}(\boldsymbol{\theta}, \boldsymbol{\kappa})\Big] + \rho_{\mathtt{s}}\Big[\mathcal{L}_{\mathtt{s}}\big(\boldsymbol{\theta}, \boldsymbol{\vartheta}, \boldsymbol{\kappa}\big) - \phi_{\mathtt{s}}(\boldsymbol{\theta}, \boldsymbol{\vartheta})\Big], \tag{10}$$

where $\rho_{\mathtt{p}}$ and $\rho_{\mathtt{s}}$ are positive penalty coefficients. The enhancer is then updated by projected gradient descent (Bertsekas, 1999):

$$\boldsymbol{\theta}^{t+1} = \Pi_{\Theta}\left(\boldsymbol{\theta}^t - \eta_{\boldsymbol{\theta}} \nabla_{\boldsymbol{\theta}} \mathcal{J}(\boldsymbol{\theta}^t)\right), \tag{11}$$

where Π_{Θ} denotes projection onto the feasible parameter space and $\eta_{\boldsymbol{\theta}}$ is the learning rate.

In this way, the lower-level optimization still follows the alternating update between $\mathcal{L}_{\mathtt{p}}$ and $\mathcal{L}_{\mathtt{s}}$, while V-PBGD converts their implicit optimality requirements into value-function penalties for efficiently updating the upper-level enhancer.

Task-Specific Adaptation (TSA). After cross-task collaborative learning, the enhancer $\boldsymbol{\theta}^*$ acquires task-aware enhancement capability through pixel-semantic collaboration. For downstream adaptation, we freeze the enhancer and fine-tune the semantic model: $\boldsymbol{\pi}^{m+1} = \boldsymbol{\pi}^m - \eta_{\boldsymbol{\pi}} \nabla_{\boldsymbol{\pi}} \ell_{\mathtt{s}}\big(\boldsymbol{\pi}^m, \boldsymbol{\theta}^*\big)$, where $\eta_{\boldsymbol{\pi}}$ denotes the learning rate.

Overall, BiCo first learns a task-aware enhancer through bilevel pixel-semantic collaboration, followed by lightweight task-specific adaptation of the semantic model. The complete training procedure is summarized in Alg. 1.

4 EXPERIMENTAL RESULTS

4.1 IMPLEMENTATION DETAILS

Network Architecture We adopt the three-layer testing architecture in (Ma et al., 2022) as the image enhancer. For downstream perception tasks, TOOD (Feng et al., 2021) and SegFormer (Xie et al., 2021) are employed for object detection and semantic segmentation, respectively.

Benchmarks and Metrics. We use the ExDark dataset (Loh & Chan, 2019) for object detection and the ACDC dataset (Sakaridis et al., 2021) for semantic segmentation. We report mAP for detection and mIoU for segmentation, and evaluate visual quality using the no-reference IQA metrics LIQE (Zhang et al., 2023), CLIP-IQA (Wang et al., 2023), and HyperIQA (Su et al., 2020).

Compared Methods. For low-light enhancement, we include ZeroDCE (Guo et al., 2020), RUAS (Liu et al., 2021b), SCI (Ma et al., 2022), PairLIE (Fu et al., 2023), LightenDiffusion (Jiang et al., 2024), ZERO-IG (Shi et al., 2024), QuadPrior (Wang et al., 2024), Di-Retinex (Sun et al., 2025b), and CWNet (Zhang et al., 2025). For fair evaluation, all enhancers use the same TOOD and SegFormer baselines as BiCo. We also compare with end-to-end low-light perception methods, including MAET (Cui et al., 2021), RIR (Sun et al., 2022), IA-YOLO (Liu et al., 2022), YOLA (Hong et al., 2024), and FRBNet (Sun et al., 2025a) for detection, and DANNet (Wu et al., 2021), GPS-GLASS (Lee et al., 2023), HRDA (Hoyer et al., 2022), SemLA (Qorbani et al., 2025), and CISS (Sakaridis et al., 2025) for segmentation.

Parameter Settings. We use a batch size of 2 for both dark object detection and semantic segmentation. For detection, we follow the standard $2\times$ MMDetection schedule and optimize with SGD using a learning rate of 1.25×10^{-3}, momentum of 0.9, and weight decay of 1×10^{-4}. For segmentation, we adopt the standard 80k MMSegmentation schedule and use AdamW with a learning rate of 8.25×10^{-5} and weight decay of 1.82×10^{-3}. During cross-task collaboration (CTC), the auxiliary lower-level networks are optimized with SGD at a learning rate of 1×10^{-2}. The VFM-guided semantic regularization L_r and VLM-guided pixel-level objective L_p are weighted by 0.05 and 0.1, respectively, while the shared penalty coefficients $\rho_p = \rho_s$ are linearly increased from 0.05 to 0.2 over the first 6k iterations. During task-specific adaptation (TSA), we initialize from the CTC-trained checkpoints and freeze the enhancement networks. We then fine-tune only the TOOD head for 5 epochs for detection and only the SegFormer head for 8k iterations for segmentation.

4.2 COMPARED WITH CASCADED FRAMEWORKS

Joint Enhancement-Detection. As shown in the second to fifth columns of Table 2, BiCo achieves the best scores across all four metrics, with particularly clear gains in LIQE and detection accuracy. The simultaneous improvements in perceptual quality and mAP indicate that the learned enhancement does not trade semantic discriminability for visual enhancement, but instead produces representations favorable to both objectives. The left part of Fig. 3 further supports this observation: competing methods may retain insufficient illumination (*e.g.*, MAET), introduce noticeable color or contrast distortion (*e.g.*, IA-YOLO and YOLA), or even yield incorrect category predictions after enhancement (*e.g.*, RIR misclassifying the dog as a cup). In contrast, BiCo recovers clearer object structures and more balanced illumination while maintaining reliable recognition and localization for both examples.

Table 2: Quantitative results on joint pixel-semantic tasks.

Method	LIQE	CLIPIQA	HyperIQA	mAP	LIQE	CLIPIQA	HyperIQA	mIoU	Params (M)
ZeroDCE	1.905	0.388	0.405	42.29	1.067	0.211	0.333	62.38	0.0794
RUAS	1.181	0.303	0.364	38.50	1.004	0.234	0.255	61.50	0.0367
SCI	1.785	0.432	0.415	42.15	1.026	0.230	0.301	60.91	**0.0003**
PairLIE	1.157	0.293	0.373	40.84	1.019	0.245	0.299	60.67	0.3418
LightenDiff	1.785	0.355	0.359	41.66	1.008	0.197	0.279	62.32	27.833
ZERO-IG	1.314	0.271	0.350	39.66	1.010	0.242	0.253	60.47	0.0866
QuadPrior	1.890	0.370	0.371	41.29	1.003	**0.294**	0.235	60.63	1313.6
Di-Retinex	1.465	0.333	0.359	41.61	1.008	0.205	0.291	61.00	0.0421
CWNet	1.854	0.369	0.402	41.48	1.012	0.211	0.318	61.37	1.4117
BiCo	**2.008**	**0.437**	**0.431**	**43.55**	**1.068**	0.248	**0.344**	**63.23**	**0.0003**

Low-Light Image Enhancement *Dark Object Detection* *Nighttime Semantic Segmentation*

Joint Enhancement-Segmentation. As shown in the sixth to ninth columns of Table 2, although BiCo ranks second on CLIPIQA, it achieves the best LIQE and HyperIQA together with the highest mIoU. Notably, QuadPrior leads only on CLIPIQA, while remaining less competitive on the other enhancement metrics and segmentation accuracy. This discrepancy suggests that a single perceptual metric may not fully reflect task-relevant enhancement quality, whereas BiCo exhibits a more

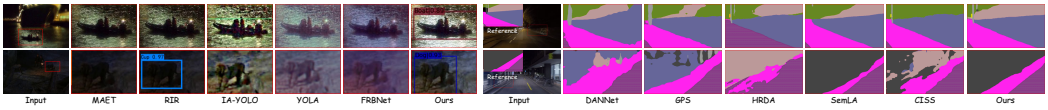

Figure 3: Qualitative comparisons on joint enhancement-detection (Left) and enhancement-segmentation (Right) tasks. Best viewed zoomed in for details.

consistent advantage across both visual quality and semantic perception. The right part of Fig. 3 further shows that competing methods often produce fragmented semantic regions (*e.g.*, GPS and HRDA), inaccurate class assignments, or irregular boundaries (*e.g.*, CISS). In contrast, BiCo yields more coherent semantic regions and cleaner class transitions that better align with the reference.

Model Efficiency. Despite using the same lightweight architecture and parameter count as SCI, the smallest among the compared enhancers, BiCo consistently improves all enhancement and downstream perception metrics, including relative gains of 3.3% in mAP and 3.8% in mIoU. This clearly indicates that the performance improvements arise from the proposed pixel-semantic collaborative learning strategy rather than increased model capacity.

4.3 Compared with End-to-End Semantic-Level Networks

Detection. *Quantitative comparison.* As shown in Table 3, BiCo achieves the best overall detection performance and remains highly competitive across most categories. In contrast, several competing methods exhibit noticeable category imbalance: IA-YOLO achieves the best results on two categories but performs worst on Car, while FRBNet leads on only three categories. BiCo achieves the best or tied-best performance on six categories, indicating more balanced detection across diverse object types. Notably, BiCo achieves the highest mAP with roughly half the parameters of the most parameter-efficient competing method, indicating that its gains arise from effective pixel-semantic collaboration rather than increased model capacity. 2) *Qualitative comparison*. The qualitative results in the left part of Fig. 4 further support this observation. Several competing methods suffer from missed detections or unstable category predictions under severe low-light degradation, with RIR even misclassifying the dog as a cup. In contrast, BiCo correctly recognizes and localizes both examples, indicating that its enhancement better preserves discriminative cues for semantic recognition.

Table 3: Comparisons with end-to-end detectors.

Method	Bicycle	Boat	Bottle	Bus	Car	Cat	Chair	Cup	Dog	Motorbike	People	Table	mAP	Params (M)
MAET	40.6	27.4	37.9	48.3	40.5	39.3	32.8	37.1	42.6	39.1	34.5	20.2	36.7	61.583
RIR	36.6	37.8	52.3	39.4	41.0	34.9	33.9	41.1	37.6	39.5	24.9	28.8	37.3	61.715
IA-YOLO	45.4	41.5	**58.3**	53.7	14.2	31.0	27.2	30.7	44.4	**51.2**	25.1	28.5	37.6	61.737
YOLA	48.0	35.1	46.1	**63.5**	48.8	41.2	33.3	37.4	48.5	40.6	38.4	**28.9**	42.5	61.592
FRBNet	**50.2**	33.7	41.4	61.0	48.3	42.1	**38.9**	**42.0**	46.7	43.6	40.4	26.7	42.9	61.592
BiCo	49.1	**38.9**	47.2	60.3	**51.3**	**44.3**	32.4	**42.0**	**47.1**	43.4	**41.4**	25.3	**43.6**	**32.044**

Segmentation. *Quantitative comparison*. As shown in Table 4, BiCo achieves the best overall segmentation performance and ranks first on 11 of the 17 semantic categories. Notably, these leading results span scene structures, traffic elements, pedestrians, and vehicles, demonstrating consistent improvements across heterogeneous semantics. In contrast, competing methods tend to favor particular categories while degrading substantially on others; for example, SemLA performs strongly on several large-area classes but poorly on Rider, while CISS achieves the best Terrain result yet struggles severely on Sky. Moreover, BiCo achieves the highest mIoU with about 40% fewer parameters than the lightest competing method, demonstrating favorable accuracy–efficiency trade-offs. 2) *Qualitative comparison*. The qualitative results in the right part of Fig. 4 show a similar trend. Competing methods often produce fragmented regions or inaccurate semantic boundaries, whereas BiCo yields more coherent predictions that better preserve scene structures and align with the reference.

Table 4: Comparison with end-to-end segmentation methods.

Method	Road	Sidewalk	Building	Wall	Fence	Pole	Traffic Light	Traffic Sign	Vegetation	Terrain	Sky	Person	Rider	Car	Train	Motorcycle	Bicycle	mIoU	Params
DANNet	91.0	61.7	71.1	32.5	36.0	20.9	37.7	22.2	69.5	21.6	80.4	32.5	16.6	42.5	61.6	8.5	46.9	44.3	74.733
GPS	92.3	67.7	68.9	33.0	32.6	37.7	46.5	31.1	68.6	20.6	81.7	21.6	10.0	53.8	11.3	5.8	36.8	42.4	74.733
HRDA	80.9	28.6	69.0	39.3	26.4	54.7	41.9	33.8	41.9	25.2	0.0	41.8	11.2	68.6	65.7	14.2	50.4	40.8	85.686
SemLA	94.2	75.5	81.7	**63.9**	**53.9**	42.8	50.3	49.7	**73.9**	11.4	**85.9**	43.8	7.2	72.7	90.8	**25.5**	56.4	57.6	455.95
CISS	95.1	75.4	72.8	38.6	39.4	55.0	54.6	41.4	43.0	**26.5**	1.8	48.7	12.0	71.6	89.9	25.1	52.4	49.6	85.686
BiCo	**96.6**	**82.0**	**82.7**	57.8	50.4	**60.5**	**69.5**	**57.2**	73.6	16.9	83.2	**59.4**	**38.8**	**77.5**	**91.2**	21.1	**56.7**	**63.2**	**44.604**



Figure 4: Qualitative comparisons with semantic-level networks for low-light object detection (Left) and semantic segmentation (Right). Best viewed zoomed in for details.

5 Algorithmic Analysis

5.1 Necessity of Bilevel Modeling

Performance Evaluation. The task-matched results in the upper-left and lower-right regions of Table 5 further verify the necessity of bilevel modeling. Under both En+De and En+Se settings, BiCo consistently improves enhancement quality together with the corresponding downstream accuracy. Compared with naive single-level learning, this alleviates the trade-off between pixel- and semantic-level objectives highlighted in Table 1, avoiding gains in one task at the expense of the other. By allowing semantic objectives to provide feedback to enhancer optimization, bilevel learning promotes coordinated improvement between the two levels. Consequently, BiCo achieves the best performance across all task-matched metrics, consistent with the landscape in Fig. 1, where collaborative feedback drives the two objectives toward a shared optimum.

Convergence Behaviors. Fig. 5 compares the optimization behaviors of BiCo and naive learning. For detection, the CTC stage consistently achieves lower loss and faster convergence, indicating that pixel-semantic collaboration facilitates optimization on the seen task. More importantly, for segmentation, both settings face an unseen downstream task and train the semantic network from scratch, making the learned enhancer the primary source of difference. Despite being learned only with detection during CTC, the BiCo enhancer leads to consistently lower and more stable segmentation loss. This further demonstrates that the learned enhancer captures transferable representations that facilitate optimization beyond the task used during collaborative learning.

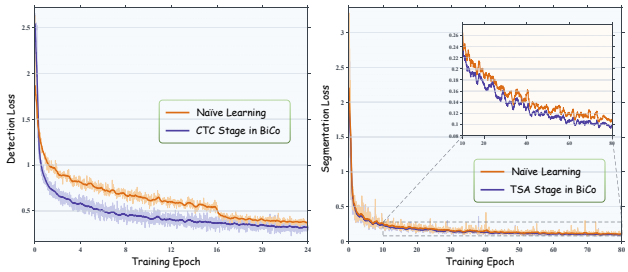

Figure 5: Convergence comparison between BiCo and naive learning. The left panel shows detection loss during CTC, while the right panel shows segmentation loss during TSA for an unseen downstream task.

Table 5: Necessity of bilevel modeling and cross-task adaptability under different task settings.

CTC Setting	Model	TSA Setting: En + De (Rows 3-4: Seen; 5-6: Unseen)				TSA Setting: En + Se (Rows 3-4: Unseen; 5-6: Seen)			
		LIQE	CLIPIQA	HyperIQA	mAP	LIQE	CLIPIQA	HyperIQA	mIoU
En + De	Naive (*Single Level*)	1.785	0.432	0.415	42.15	1.026	**0.230**	0.301	60.91
	BiCo (*Bilevel*)	**2.008**↑0.223	**0.437**↑0.005	**0.431**↑0.016	**43.55**↑1.40	**1.037**↑0.011	0.199↓0.031	**0.324**↑0.023	**62.69**↑1.78
En + Se	Naive (*Single Level*)	**1.785**	**0.432**	**0.415**	42.15	1.026	0.230	0.301	60.91
	BiCo (*Bilevel*)	1.515↓0.270	0.330↓0.102	**0.415**0.000	**42.36**↑0.21	**1.068**↑0.042	**0.248**↑0.018	**0.344**↑0.043	**63.23**↑2.32

5.2 Cross-Task Adaptability

As evidenced by the upper-right and lower-left regions of Table 5, BiCo exhibits consistent cross-task gains at the semantic level: the enhancer learned under En + De still improves unseen segmentation, while that learned under En + Se also benefits detection. Interestingly, the enhancement metrics show different transfer behaviors. The En + De enhancer largely preserves or improves perceptual quality when transferred to segmentation, whereas the En + Se enhancer shows larger variations in IQA scores despite still improving detection accuracy. Nevertheless, the consistent downstream gains indicate that BiCo learns task-agnostic enhancement representations that preserve semantic information transferable across tasks.

5.3 NECESSITY OF JOINT PIXEL-SEMANTIC CONSTRAINTS

Table 6: Necessity of joint pixel-semantic constraints.

$\mathcal{L}_{\mathrm{p}}$	$\mathcal{L}_{\mathrm{s}}$	LIQE	CLIPIQA	HyperIQA	mAP
✓	✗	1.916	0.426	0.424	42.16
✗	✓	1.960	0.434	0.428	42.82
✓	✓	**2.008**	**0.437**	**0.431**	**43.55**

Table 6 investigates the roles of pixel- and semantic-level constraints. Using either $\mathcal{L}_{\mathrm{p}}$ or $\mathcal{L}_{\mathrm{s}}$ alone mainly favors its corresponding objective, leading to limited or unbalanced gains between enhancement quality and downstream perception. In contrast, jointly introducing both constraints consistently achieves the best performance across all perceptual-quality and detection metrics, indicating complementary rather than competing supervision. Specifically, the pixel-level constraint promotes visually favorable enhancement, while the semantic-level constraint helps preserve task-relevant structures and discriminative cues. Their joint use therefore provides more balanced guidance for pixel-semantic collaborative learning.

5.4 INTERMEDIATE RESULTS

Table 7: Intermediate results.

Method	LIQE	CLIPIQA	HyperIQA	mAP	LIQE	CLIPIQA	HyperIQA	mIoU
Base	1.785	0.432	0.415	42.15	1.026	0.230	0.301	60.91
+ CTC	2.008	0.437	0.431	43.18	1.068	0.248	0.344	62.55
+ CTC + TSA	**2.008**	**0.437**	**0.431**	**43.55**	**1.068**	**0.248**	**0.344**	**63.23**

Table 7 analyzes the intermediate results of the two-stage optimization. After CTC, both enhancement quality and downstream perception performance are consistently improved, indicating that joint pixel-semantic collaboration provides effective semantic feedback for optimizing the enhancer while simultaneously benefiting the semantic model. This confirms that the first stage already establishes a mutually beneficial interaction between enhancement and perception. Subsequently, TSA keeps the learned enhancer fixed, as reflected by the unchanged enhancement metrics, and further improves the downstream performance in terms of both mAP and mIoU. The results therefore demonstrate that CTC is responsible for learning transferable enhancement representations, while TSA further specializes the semantic model to the target task.

5.5 EXPLORING ENHANCER META-REPRESENTATION

Table 8: Model recombination.

Setting		LIQE	CLIPIQA	HyperIQA	mAP/mIoU
Base	En + De	1.785	0.432	0.415	42.15
	En + Se	1.026	0.230	0.301	60.91
CTC	En + De	2.008	0.437	0.431	43.55
	En + Se	1.068	0.248	0.344	63.23
Recom	En + De	1.515	0.330	0.415	42.98
	En + Se	1.037	0.199	0.324	62.99

Table 8 examines the meta-representation capability of the learned enhancer through model recombination. Specifically, we swap the enhancers learned under detection- and segmentation-oriented CTC settings and combine them with the semantic network from the other task. Although the recombined models involve mismatched task settings, both still outperform their corresponding Base models and remain close to the task-matched CTC results. This suggests that the learned enhancer is not strongly tied to a specific semantic task, but instead captures task-agnostic representations that can be transferred across downstream tasks. The recombination results therefore provide further evidence of the enhancer's meta-level generalization beyond the task used during collaborative learning.

6 CONCLUSION

We propose BiCo, a bilevel collaborative framework for jointly optimizing enhancement and semantic perception in low-light vision. By explicitly modeling their hierarchical dependency, BiCo enables reciprocal feedback between pixel- and semantic-level objectives and alleviates the trade-off caused by conventional single-level optimization. With role-balanced evaluators and two-stage learning, BiCo learns a transferable enhancer that generalizes across both seen and unseen tasks. Extensive experiments demonstrate consistent improvements in visual quality, task performance, and cross-task adaptability.

Future work will investigate whether the learned meta-representation can generalize to broader degradation conditions, more diverse downstream tasks, and more complex real-world environments. It would also be valuable to explore how pixel-semantic collaboration can be extended beyond single-image low-light vision to temporally varying or multimodal settings, further improving the scalability and generalization of the proposed learning paradigm.

AI USE STATEMENT

In this work, we used generative AI tools only for language editing and polishing. We did not use generative AI tools for generating references, developing research ideas, designing methodologies or experiments, or analyzing results. All AI-assisted edits were reviewed by the authors. We take full responsibility for the final content of this work.

REFERENCES

Dimitri P. Bertsekas. *Nonlinear Programming*. Athena Scientific, 1999.

Alexandru Brateanu, Tingting Mu, Codruta O Ancuti, and Cosmin Ancuti. Multinex: Lightweight low-light image enhancement via multi-prior retinex. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 29887–29896, 2026.

Haopeng Chen, Yihao Ai, Kabeen Kim, Robby T Tan, Yixin Chen, and Bo Wang. Udapose: Unsupervised domain adaptation for low-light human pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13781–13792, 2026.

Joshua Cho, Sara Aghajanzadeh, Zhen Zhu, and David Forsyth. Mr. illuminate: Zero-shot low-light image enhancement with diffusion prior. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8460–8470, 2026.

Ziteng Cui, Guo-Jun Qi, Lin Gu, Shaodi You, Zenghui Zhang, and Tatsuya Harada. Multitask aet with orthogonal tangent regularity for dark object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 2553–2562, 2021.

Dengxin Dai and Luc Van Gool. Dark model adaptation: Semantic image segmentation from daytime to nighttime. In *ITSC*, pp. 3819–3824, 2018.

Zhipeng Du, Miaojing Shi, and Jiankang Deng. Boosting object detection with zero-shot day-night domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12666–12676, June 2024.

Chengjian Feng, Yujie Zhong, Yu Gao, Matthew R. Scott, and Weilin Huang. Tood: Task-aligned one-stage object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 3510–3519, 2021.

Zhenqi Fu, Yan Yang, Xiaotong Tu, Yue Huang, Xinghao Ding, and Kai-Kuang Ma. Learning a simple low-light image enhancer from paired low-light instances. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22252–22261, 2023.

Chunle Guo, Chongyi Li, Jichang Guo, Chen Change Loy, Junhui Hou, Sam Kwong, and Runmin Cong. Zero-reference deep curve estimation for low-light image enhancement. In *CVPR*, pp. 1780–1789, 2020.

Khurram Azeem Hashmi, Goutham Kallempudi, Didier Stricker, and Muhammad Zeshan Afzal. Featenhancer: Enhancing hierarchical features for object detection and beyond under low-light vision. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 6725–6735, 2023.

Mingbo Hong, Shen Cheng, Haibin Huang, Haoqiang Fan, and Shuaicheng Liu. You only look around: Learning illumination-invariant feature for low-light object detection. *Advances in Neural Information Processing Systems*, 37:87136–87158, 2024.

Lukas Hoyer, Dengxin Dai, and Luc Van Gool. Hrda: Context-aware high-resolution domain-adaptive semantic segmentation. In *European Conference on Computer Vision*, pp. 372–391, 2022.

Zeyu Hua, Hui Li, Yu Wang, Song Wang, Congchao Zhu, and Caixia Zheng. Bi-bridge: Bidirectional diffusion bridges for low-light image enhancement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 37455–37464, 2026.

Haofeng Huang, Yifan Li, Wenjing Wang, Wenhan Yang, Ling-Yu Duan, and Jiaying Liu. Quad-Prior++: Multi-dimension augmented physical prior for zero-reference illumination enhancement. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 48(3):3383–3398, 2026.

Hai Jiang, Ao Luo, Xiaohong Liu, Songchen Han, and Shuaicheng Liu. Lightendiffusion: Unsupervised low-light image enhancement with latent-retinex diffusion models. In *European Conference on Computer Vision*, pp. 161–179, 2024.

Yifan Jiang, Xinyu Gong, Ding Liu, Yu Cheng, Chen Fang, Xiaohui Shen, Jianchao Yang, Pan Zhou, and Zhangyang Wang. Enlightengan: Deep light enhancement without paired supervision. *IEEE TIP*, 30:2340–2349, 2021.

Edwin H Land and John J McCann. Lightness and retinex theory. *Journal of the Optical Society of America*, 1971.

Hongjae Lee, Changwoo Han, Jun-Sang Yoo, and Seung-Won Jung. Gps-glass: Learning nighttime semantic segmentation using daytime video and gps data. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4001–4010, 2023.

Attila Lengyel, Sourav Garg, Michael Milford, and Jan C van Gemert. Zero-shot day-night domain adaptation with a physics prior. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4399–4409, 2021.

Yifan Li, Haofeng Huang, Wenhan Yang, and Jiaying Liu. Towards generalized representations for low-light understanding: When signal constancy meets semantic enrichment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1386–1395, 2026.

Dong Liang, Ling Li, Mingqiang Wei, Shuo Yang, Liyan Zhang, Wenhan Yang, Yun Du, and Huiyu Zhou. Semantically contrastive learning for low-light image enhancement. In *AAAI*, 2022.

Jinxiu Liang, Jingwen Wang, Yuhui Quan, Tianyi Chen, Jiaying Liu, Haibin Ling, and Yong Xu. Recurrent exposure generation for low-light face detection. *IEEE Transactions on Multimedia*, 24:1609–1621, 2021.

Risheng Liu, Jiaxin Gao, Jin Zhang, Deyu Meng, and Zhouchen Lin. Investigating bi-level optimization for learning and vision from a unified perspective: A survey and beyond. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(12):10045–10067, 2021a.

Risheng Liu, Long Ma, Jiaao Zhang, Xin Fan, and Zhongxuan Luo. Retinex-inspired unrolling with cooperative prior architecture search for low-light image enhancement. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10561–10570, 2021b.

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.

Yuen Peng Loh and Chee Seng Chan. Getting to know low-light images with the exclusively dark dataset. *Computer Vision and Image Understanding*, 178:30–42, 2019.

Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *CVPR*, pp. 3431–3440, 2015.

Long Ma, Tengyu Ma, Risheng Liu, Xin Fan, and Zhongxuan Luo. Toward fast, flexible, and robust low-light image enhancement. In *CVPR*, pp. 5637–5646, 2022.

Reza Qorbani, Gianluca Villani, Theodoros Panagiotakopoulos, Marc Botet Colomer, Linus Härenstam-Nielsen, Mattia Segu, Pier Luigi Dovesi, Jussi Karlgren, Daniel Cremers, Federico Tombari, and Matteo Poggi. Semantic library adaptation: Lora retrieval and fusion for open-vocabulary semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9804–9815, 2025.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, 2021.

Christos Sakaridis, Dengxin Dai, and Luc Van Gool. Guided curriculum model adaptation and uncertainty-aware evaluation for semantic nighttime image segmentation. In *ICCV*, pp. 7374–7383, 2019.

Christos Sakaridis, Dengxin Dai, and Luc Van Gool. Map-guided curriculum domain adaptation and uncertainty-aware evaluation for semantic nighttime image segmentation. *IEEE TPAMI*, 2020.

Christos Sakaridis, Dengxin Dai, and Luc Van Gool. Acdc: The adverse conditions dataset with correspondences for semantic driving scene understanding. In *ICCV*, pp. 10765–10775, 2021.

Christos Sakaridis, David Bruggemann, Fisher Yu, and Luc Van Gool. Condition-invariant semantic segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(4):3111–3125, 2025.

Han Shen and Tianyi Chen. On penalty-based bilevel gradient descent method. In *International Conference on Machine Learning*, pp. 30992–31015, 2023.

Yiqi Shi, Duo Liu, Liguo Zhang, Ye Tian, Xuezhi Xia, and Xiaojing Fu. Zero-ig: Zero-shot illumination-guided joint denoising and adaptive enhancement for low-light images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3015–3024, 2024.

Gwangik Shin, Jaeha Song, and Soonmin Hwang. Self-guided low light object detection framework. In *The Fourteenth International Conference on Learning Representations*, 2026.

Oriane Siméoni, Huy V Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, et al. Dinov3. *arXiv preprint arXiv:2508.10104*, 2025.

Shaolin Su, Qingsen Yan, Yu Zhu, Cheng Zhang, Xin Ge, Jinqiu Sun, and Yanning Zhang. Blindly assess image quality in the wild guided by a self-adaptive hyper network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3667–3676, 2020.

Fangtong Sun, Congyu Li, Ke Yang, Yuchen Pan, Hanwen Yu, Xichuan Zhang, and Yiying Li. Frbnet: Revisiting low-light vision through frequency-domain radial basis network. *Advances in Neural Information Processing Systems*, 38:99085–99113, 2025a.

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.

Shangquan Sun, Wenqi Ren, Jingyang Peng, Fenglong Song, and Xiaochun Cao. Di-retinex: Digital-imaging retinex model for low-light image enhancement. *International Journal of Computer Vision*, pp. 1–22, 2025b.

Jianyi Wang, Kelvin C. K. Chan, and Chen Change Loy. Exploring clip for assessing the look and feel of images. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 2555–2563, 2023.

Wenjing Wang, Wenhan Yang, and Jiaying Liu. Hla-face: Joint high-low adaptation for low light face detection. In *CVPR*, pp. 16195–16204, 2021.

Wenjing Wang, Huan Yang, Jianlong Fu, and Jiaying Liu. Zero-reference low-light enhancement via physical quadruple priors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 26057–26066, 2024.

Xinyi Wu, Zhenyao Wu, Hao Guo, Lili Ju, and Song Wang. Dannet: A one-stage domain adaptation network for unsupervised nighttime semantic segmentation. In *CVPR*, pp. 15769–15778, 2021.

Yuhui Wu, Chen Pan, Guoqing Wang, Yang Yang, Jiwei Wei, Chongyi Li, and Heng Tao Shen. Learning semantic-aware knowledge guidance for low-light image enhancement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1662–1671, 2023.

Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M. Alvarez, and Ping Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. In *Advances in Neural Information Processing Systems*, volume 34, pp. 12077–12090, 2021.

Xinwei Xue, Jia He, Long Ma, Yi Wang, Xin Fan, and Risheng Liu. Best of both worlds: See and understand clearly in the dark. In *Proceedings of the 30th ACM International Conference on Multimedia*, pp. 2154–2162, 2022.

Tongshun Zhang, Pingping Liu, Yubing Lu, Mengen Cai, Zijian Zhang, Zhe Zhang, and Qiuzhan Zhou. Cwnet: Causal wavelet network for low-light image enhancement. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 8789–8799, 2025.

Weixia Zhang, Guangtao Zhai, Ying Wei, Xiaokang Yang, and Kede Ma. Blind image quality assessment via vision-language correspondence: A multitask learning perspective. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14071–14081, 2023.

Yin Zhang, Yongqiang Zhang, Zian Zhang, Man Zhang, Rui Tian, and Mingli Ding. Isp-teacher: Image signal process with disentanglement regularization for unsupervised domain adaptive dark object detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 7387–7395, 2024.

Qing Zhao, Weijian Deng, Pengxu Wei, ZiYi Dong, Hannan Lu, Xiangyang Ji, and Liang Lin. Delving into cascaded instability: a lipschitz continuity view on image restoration and object detection synergy. *Advances in Neural Information Processing Systems*, 38:164856–164881, 2025.

Naishan Zheng, Jie Huang, Man Zhou, Zizheng Yang, Qi Zhu, and Feng Zhao. Learning semantic degradation-aware guidance for recognition-driven unsupervised low-light image enhancement. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 3678–3686, 2023.