DLDA: Unified Dual-Level Domain Adaptation for Low-Light Object Detection

Jiayi Hu[1, 4], **Qian Zhao**[3, 4], **Gang Li**[1, 2, 3, 4*]
[1]College of Electronics and Information Engineering, Tongji University
[2]Shanghai Institute of Intelligent Science and Technology, Tongji University
[3]Shanghai Research Institute for Intelligent Autonomous Systems, Tongji University
[4]National Key Laboratory of Autonomous Intelligent Unmanned Systems, China
yiyiyi@tongji.edu.cn, qianzhao@tongji.edu.cn, lig@tongji.edu.cn

Abstract

Low-light object detection faces significant challenges due to the substantial domain shift between normal-light and low-light conditions. Prior works often enhance low-light images before detection, but this preprocessing can introduce artifacts that degrade detection performance since it focuses on human visual quality rather than task-specific features. Other methods incorporate illumination-aware modules for low-light feature learning, yet their scalability is limited by the scarcity of annotated low-light datasets. To overcome these limitations, we propose a unified Dual-Level Domain Adaptation (DLDA) framework that jointly addresses image-level and feature-level domain discrepancies. Specifically, we introduce a luminance-aware contrastive translation module that synthesizes target-style low-light images while preserving structural details, enabling effective image-level adaptation. Building on this, we further design a multi-scale conditional adversarial alignment strategy that promotes semantic consistency across feature hierarchies to enhance domain-invariant feature extraction. Extensive experiments on multiple low-light detection benchmarks demonstrate that DLDA achieves state-of-the-art performance, exhibiting strong robustness and generalization.

Code — https://github.com/starsclouds/DLDA

Introduction

Object detection, which aims to identify and localize objects in images, is a fundamental problem in computer vision with applications in autonomous driving, video surveillance, and robotics (Li et al. 2023; Zhang et al. 2016; Hong et al. 2021). With the rapid development of deep convolutional neural networks and large-scale datasets such as COCO (Lin et al. 2014) and Pascal VOC (Everingham et al. 2010), significant progress has been achieved under well-lit conditions. However, in real-world low-light scenarios, images often suffer from low contrast, noise, and texture loss, causing a severe domain gap between normal-light and low-light images and leading to substantial performance degradation (Yang et al. 2020; Wu et al. 2023; Zhao et al. 2025; Li et al. 2025).

A common solution is to enhance low-light images before detection (Liang et al. 2023; Wang et al. 2024; Yan et al. 2024; Yin et al. 2023). While these methods improve visual quality, they are designed for human perception rather than machine perception and may introduce artifacts that degrade detection performance (Yang et al. 2023; Huang, Le, and Jaw 2020). More recent approaches directly modify detection architectures by embedding illumination-aware modules to enhance feature representation. Representative methods include MAET (Cui et al. 2021), which employs multi-task learning with multi-exposure images to enhance robustness, while IAT (Cui et al. 2022) leverages a lightweight transformer for parameter-efficient enhancement. To address the lack of hierarchical context in dark scenes, FeatEnHancer (Hashmi et al. 2023) reinforces multi-scale features for richer contextual representation. Furthermore, YOLA (Hong et al. 2024) introduces illumination-invariant learning to reduce lighting-specific bias. Despite these advances, most of these methods (Zhang, Suo, and Dai 2023; Ma et al. 2022; Liu et al. 2022) rely on scarce low-light datasets with annotations, which are difficult to obtain. Moreover, few approaches exploit abundant normal-light data for cross-domain adaptation to bridge the low-light detection gap. A notable attempt is DAI (Du, Shi, and Deng 2024), which introduces illumination-invariant cues to mitigate image-level appearance shifts, yet it overlooks discrepancies in the feature domain distributions.

To address these limitations, inspired by DAI, we propose **DLDA (Unified Dual-Level Domain Adaptation)**, as shown in Fig. 1, a framework that unifies image-level and feature-level adaptation. Within this framework, we design two complementary modules: one addressing image-level style discrepancies through image-level adaptation, and the other achieving feature alignment across multiple feature scales. By jointly considering these two levels, DLDA effectively handles both appearance and feature domain gaps and achieves state-of-the-art performance on challenging low-light detection benchmarks.

Our contributions are as follows:

- We propose a unified dual-level domain adaptation framework that jointly performs image-level style translation and feature-level alignment, effectively bridging the domain gap between normal-light and low-light images for object detection. Extensive experiments on multiple low-light detection benchmarks demonstrate that DLDA outperforms existing state-of-the-art methods.

[*]Corresponding author.

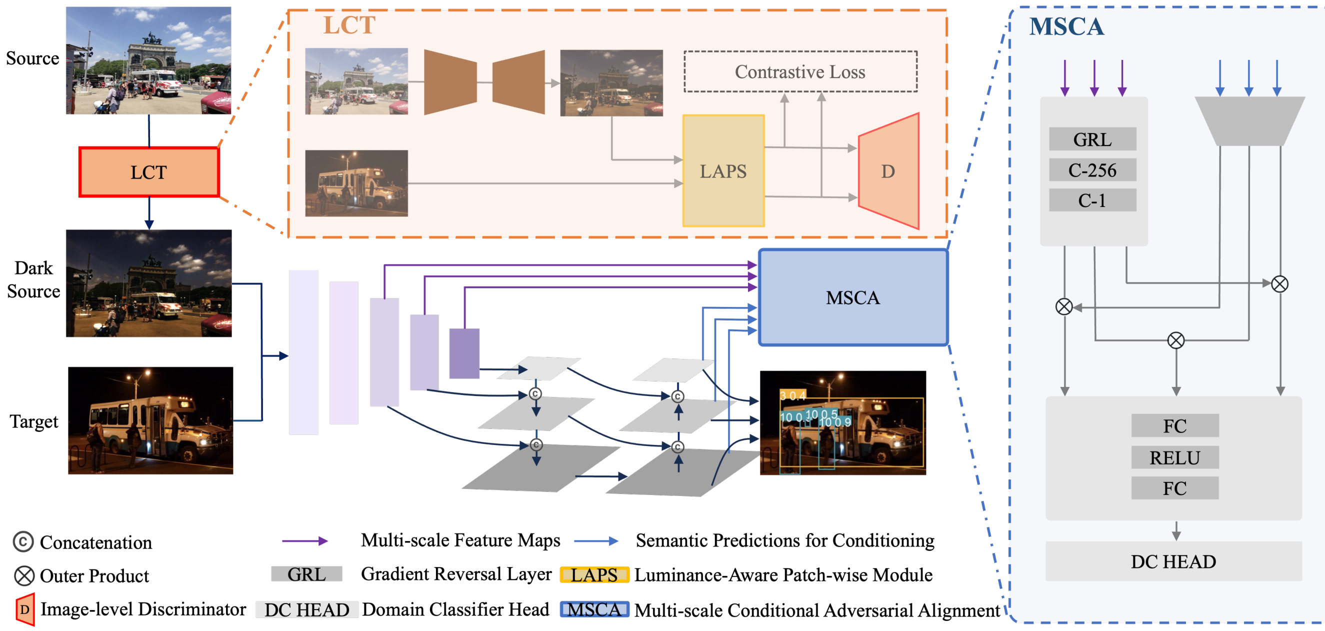


Figure 1: Overview of the proposed DLDA framework. Our contributions include: (1) We generate low-light images from our LCT module, and feed them into subsequent framework together with unpaired target domain images. (2) An assisted auxiliary feature-level domain adaptation branch(MSCA). During the training stage, the multi-level features extracted from the backbone network and the categorical features from the detection head are jointly fed into this branch to enhance domain-invariant feature learning. During the test stage, this branch is removed to accelerate the inference speed.

- We introduce a luminance-aware contrastive image translation approach for low-light scenarios, which effectively aligns the styles of source and target domains in the pixel space and provides more discriminative inputs for detection tasks.
- We design a multi-scale conditioned feature alignment strategy that achieves fine-grained domain consistency in the feature space, significantly enhancing the robustness and generalization of the model for cross-domain detection under low-light conditions.

Related Work

Object Detection

Modern object detectors can be broadly categorized into two-stage and one-stage approaches. In two-stage detectors, such as Faster R-CNN (Ren et al. 2015) and Mask R-CNN (He et al. 2017), region proposals are first generated and then followed by classification and bounding-box regression. One-stage detectors, represented by the YOLO series (Jocher et al. 2021) and RetinaNet (Lin et al. 2017), directly predict class probabilities and bounding boxes on dense feature maps. While slightly less accurate, one-stage designs are simpler and more efficient, making them attractive for real-time and resource-constrained scenarios. Recently, Transformer-based detectors, including DETR (Carion et al. 2020) and its variants (Zhang et al. 2023), have shown strong performance but typically require larger models and longer training. However, most detectors trained on well-lit data suffer significant performance drops in low-light or adverse conditions. In this work, we adopt YOLOv5-l as our baseline due to its balance of accuracy and efficiency.

Low-light Object Detection

Existing methods for low-light object detection can be broadly categorized into two lines of research: low-light image enhancement and low-light feature learning. Low-light image enhancement aims to improve the perceptual quality of input images prior to detection. Representative approaches include the deep decomposition–enhancement framework KIND (Zhang et al. 2019), the multi-scale fusion-based MBLLEN (Lv et al. 2018), and the zero-reference curve estimation model Zero-DCE (Guo et al. 2020). While these enhancement techniques significantly improve visual quality for human perception, they often introduce artifacts that can impair downstream detection performance.

Recent studies focuse on designing detectors that directly learn robust representations under low-light conditions. Notable examples include FeatEnHancer (Hashmi et al. 2023), which reinforces multi-scale features, IAT (Cui et al. 2022), which incorporates illumination-aware transformer modules, MAET (Cui et al. 2021), which integrates a multi-task autoencoder mechanism, and YOLA (Hong et al. 2024), which adopts illumination-invariant feature learning. These methods are designed to extract discriminative representations from low-light inputs that are better aligned with detection objectives, thereby enhancing performance under challenging illumination conditions.

Despite notable progress, current methods are limited by scarce low-light annotations and the absence of a unified framework that combines image-level and feature-level modeling. We propose a unified design integrating image-level translation and feature-level adaptation to better exploit normal-light data and enhance low-light detection.

Domain Adaptation Strategies

Domain adaptation has emerged as an effective paradigm to mitigate domain shift without requiring extensive annotations in the target domain. Although it has been widely studied in generic object detection, its application to low-light scenarios has received limited attention. Representative attempts include synthesizing low-light images to reduce image-level discrepancies (Cui et al. 2021; Yang and Soatto 2020; Arruda et al. 2019; Luo et al. 2023; Bhattacharjee et al. 2020; Lin et al. 2023). Another line of research focuses on feature-level alignment by learning domain-invariant representations. Early studies employ adversarial learning to align source and target feature distributions (Ganin et al. 2016; Long et al. 2018), while recent advances have enhanced alignment through improved discriminators. For example, DAYOLO (Li et al. 2022) introduces adversarial learning to bridge day–night features, MS-DAYOLO (Hnewa and Radha 2021) extends this idea with multi-scale discriminators, and DAI (Du, Shi, and Deng 2024) incorporates illumination-invariant cues into the detection head to improve robustness under low-light conditions.

Despite these advances, most existing methods focus either on image-level style translation or feature-level representation learning. In practice, large domain gaps caused by illumination, sensor, and viewpoint variations make it challenging to extract truly invariant features. To address this issue, we propose a unified framework that jointly leverages image-level and feature-level adaptation, effectively bridging the gap between normal- and low-light domains for robust detection.

Method

Overall Framework

We aim to adapt object detectors trained on well-lit images to unlabeled low-light domains, formulated as an unsupervised domain adaptation task. In this setting, labeled data are available only in the source domain with sufficient illumination, while the target domain contains only unlabeled low-light images. The core challenge lies in the severe domain gap caused by drastic luminance degradation and structural distortion under low-light conditions, which hinders the generalization of features learned in the source domain.

To address this issue, we propose DLDA, a unified Dual-Level Domain Adaptation framework that jointly performs image-level style translation and feature-level semantic alignment, as illustrated in Fig. 1. Unlike existing methods that focus solely on either image translation or feature alignment (Du, Shi, and Deng 2024; Cui et al. 2022), DLDA integrates both levels into a cohesive architecture to learn domain-invariant and discriminative features under adverse illumination. Our approach employs a luminance-aware contrastive translation module that transforms well-lit source images into target-style low-light counterparts while preserving structural details. To prevent artifacts that may harm downstream detection, we introduce a patch-wise contrastive constraint guided by luminance, ensuring semantic consistency during translation. To further bridge feature-level domain gaps, we design a multi-scale conditional adversarial alignment strategy. By conditioning alignment on predicted class distributions, this module facilitates the learning of semantically consistent and domain-invariant representations across multiple feature hierarchies.

Luminance-aware Contrastive Translation (LCT)

To reduce image-level domain discrepancies between well-lit source images and low-light target images, we propose a *Luminance-aware Contrastive Translation (LCT)* module. Inspired by CUT (Park et al. 2020), our LCT module performs unpaired one-sided translation to synthesize structurally preserved low-light images without relying on heuristic enhancement or style transfer, thus facilitating robust detection.

Adversarial Translation with Structure Preservation. Given a source image $I_s \sim p_s$, the generator G produces a translated image $\hat{I}_t = G(I_s)$ aligned with the target low-light domain p_t. A discriminator D is trained to distinguish real target images $I_t \sim p_t$ from generated images $\hat{I}_t$. The adversarial loss is defined as:

$$\mathcal{L}_{\text{GAN}} = \mathbb{E}_{I_t}[\log D(I_t)] + \mathbb{E}_{I_s}[\log(1 - D(G(I_s)))], \tag{1}$$

Luminance-aware Patch-wise Contrastive Learning. To preserve local semantics during translation, we adopt a *luminance-aware PatchNCE* loss, which treats spatially aligned patches in I_s and $\hat{I}_t$ as positive pairs, with N negatives sampled from different locations.

To improve robustness in low-light scenes, we introduce a *contrast-enhanced luminance mask* that emphasizes semantically informative patches. A patch z is deemed valid if its luminance deviates sufficiently from its local neighborhood:

$$|\mathrm{Lum}(z) - \mu_{\mathcal{N}}(z)| > \tau_C, \tag{2}$$

where $\mathrm{Lum}(z)$ is the luminance of z, $\mu_{\mathcal{N}}(z)$ is the average luminance of its local window, and τ_C is a contrast threshold.

Let z and $\hat{z}$ denote the projected features of the source and translated patches, and $\{z_n^-\}$ be the negatives. The patch-wise contrastive loss is:

$$\mathcal{L}_{\text{PatchNCE}} = -\log \frac{q^+}{q^+ + \sum_{n=1}^{N} q_n^-}, \tag{3}$$

where $q^+ = \exp(\hat{z} \cdot z/\tau)$ and $q_n^- = \exp(\hat{z} \cdot z_n^-/\tau)$ are the similarity scores, and τ is a temperature parameter.

The final PatchNCE loss aggregates over layers and valid spatial locations using the luminance mask M_l^s:

$$\mathcal{L}^s_{\text{PatchNCE}} = \sum_{l=1}^{L} \sum_{s \in \mathcal{S}_l} M_l^s \cdot \mathcal{L}_{\text{PatchNCE}}(\hat{z}_l^s, z_l^s, \{z_l^{s_n}\}), \tag{4}$$

A similar loss $\mathcal{L}^t_{\text{PatchNCE}}$ is computed for the target domain.

Figure 2: Qualitative comparison of image-level translation methods. Our LCT module generates more realistic and semantically faithful low-light images.

Overall LCT Objective. The final LCT objective integrates adversarial and contrastive losses from both domains:

$$\mathcal{L}_{\text{LCT}} = \mathcal{L}_{\text{GAN}} + \lambda_X \mathcal{L}^{s}_{\text{PatchNCE}} + \lambda_Y \mathcal{L}^{t}_{\text{PatchNCE}}, \tag{5}$$

where λ_X and λ_Y control the contribution of contrastive losses from source and target domains, respectively. We follow standard practice and fix $\lambda_X = \lambda_Y = 1$ throughout all experiments.

Qualitative comparisons of LCT with prior translation methods are shown in Fig. 2. Compared to degradation-based DISP (Cui et al. 2021), de-enhancement-based Zero-DCE (Guo et al. 2020), and domain adaptation-based CUT (Park et al. 2020) and FDA (Yang and Soatto 2020) approaches, our LCT produces more realistic low-light images with preserved semantics, effectively supporting downstream detection tasks.

Multi-scale Conditional Adversarial Alignment (MSCA)

While LCT reduces image-level discrepancies, domain shifts in the feature space persist due to illumination-induced degradation and domain-specific semantics. To address this, we propose a *Multi-scale Conditional Adversarial Alignment (MSCA)* strategy, which performs class-conditioned adversarial learning across multiple semantic layers of the detector. Modern single-stage detectors (e.g., YOLO (Jocher et al. 2021)) extract hierarchical features $f_j = F_j(x)$ at scales $j \in \{P_3, P_4, P_5\}$, where lower-level features capture textures and higher-level ones encode semantics. However, due to illumination bias, the joint distributions $P(f_j^s, y^s)$ and $Q(f_j^t, y^t)$ can diverge significantly across domains, hindering generalization (Ganin et al. 2016). MSCA addresses this by enforcing semantic alignment at each scale, enabling the extraction of domain-invariant representations.

Conditional Adversarial Alignment Across Scales. To mitigate feature-level domain gaps, we introduce a set of discriminators $\{D_j\}$, each conditioned on class predictions p from the detection head. For each scale, we construct a joint representation via outer product:

$$h_j = f_j \otimes p, \tag{6}$$

where $\otimes$ denotes outer product, encoding interactions between spatial features and semantic predictions, following the CDAN principle (Long et al. 2018).

Each D_j aims to distinguish whether h_j originates from the source domain, using the standard adversarial loss:

$$\mathcal{L}_j^{\text{dom}} = -\mathbb{E}_s\left[\log D_j(h_j^s)\right] - \mathbb{E}_t\left[\log\left(1 - D_j(h_j^t)\right)\right], \tag{7}$$

where h_j^s and h_j^t are source and target joint representations.

Overall MSCA Objective. The total alignment loss sums over all scales:

$$\mathcal{L}_{\text{MSCA}} = \sum_{j \in \{P_3, P_4, P_5\}} \lambda_j \mathcal{L}_j^{\text{dom}}, \tag{8}$$

where λ_j controls the contribution of each scale. Training is formulated as a minimax game:

$$\min_{\theta_F, \theta_G} \left[\mathcal{L}_{\text{det}} + \lambda \cdot \mathcal{L}_{\text{MSCA}}\right], \quad \max_{\theta_D} \mathcal{L}_{\text{MSCA}}, \tag{9}$$

where θ_F, θ_G denote the feature extractor and detector, and θ_D the discriminators.

Training Strategy

Objective Functions. Our training objective combines two complementary objectives:
(1) Detection loss. We adopt the standard single-stage detection loss on the labeled source domain:

$$\mathcal{L}_{\text{det}} = \mathcal{L}_{\text{cls}} + \mathcal{L}_{\text{reg}}, \tag{10}$$

where $\mathcal{L}$cls and $\mathcal{L}_{\text{reg}}$ denote classification and bounding-box regression losses, respectively.
(2) Overall training objective. The total objective integrates detection and domain adaptation losses:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{det}} + \alpha\, \mathcal{L}_{\text{MSCA}}, \tag{11}$$

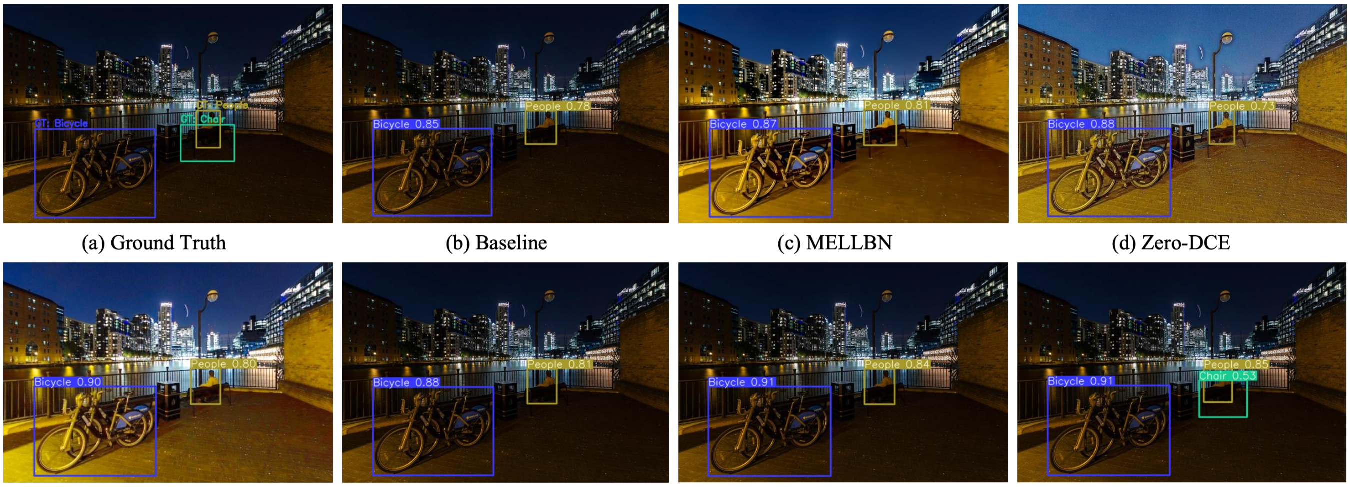
where α balances the contribution of the alignment loss and is fixed to 1 in all experiments.

Optimization and Training Schedule. To ensure stable convergence, we adopt a two-phase training strategy. In the first stage, the LCT module is optimized independently while keeping the detector frozen, allowing the generator to synthesize high-fidelity low-light translations aligned with the target domain distribution. In the second phase, the translated images are used to train the detector and the MSCA module. The LCT module is frozen, and feature-level adaptation is performed by updating the detection backbone and multi-scale discriminators according to the total loss in Eq. 11. Our model effectively leverages complementary cues from both domains, this leads to robust and transferable representations for reliable detection in the low-light target domain.

Experiments

Experimental Setup

Datasets and Tasks. We evaluate the proposed DLDA framework on two representative cross-domain object detection tasks under low-light conditions. The first task involves adapting from the well-lit COCO dataset (Lin et al. 2014) to the low-light ExDark dataset (Loh and Chan 2019). To



(a) Ground Truth (b) Baseline (c) MELLBN (d) Zero-DCE

(e) IAT (f) MAET (g) DAI (h) Ours

Figure 3: Qualitative detection results on the ExDark test set (COCO → ExDark). Our DLDA model detects more accurate and complete objects under low-light and noise conditions, outperforming baseline and state-of-the-art methods in terms of robustness and localization precision.

ensure semantic consistency, we construct a filtered version of COCO by retaining only the 12 object categories that are shared with ExDark. The ExDark dataset comprises 5,884 training images, 741 validation images, and 738 test images, covering a variety of challenging nighttime scenes. All experiments are conducted on the 12 common categories between the two domains.

The second task assesses the cross-domain generalization ability of DLDA on face detection, transferring from WIDER FACE (Yang et al. 2016) to DARK FACE (Fan et al. 2017). This setting is highly challenging due to extreme illumination degradation and frequent occlusion in the target domain.

Training Details. For image-level adaptation, both the input source images and their synthesized low-light counterparts are resized to 640×640 to maintain compatibility with the detection pipeline. During training, multi-scale features extracted from both synthetic low-light images and real target domain images are simultaneously forwarded to: (a) the standard YOLO detection head, which computes supervised detection loss on the source domain; (b) the MSCA branch, which is only active during training and is removed at inference time to ensure that no additional computational overhead is introduced. The entire framework is built upon the YOLOv5-L backbone as described earlier.

We train all models using Stochastic Gradient Descent (SGD) with a momentum of 0.937, weight decay of 0.0005, and an initial learning rate of 0.01. The batch size is set to 16, and training is conducted for a total of 300 epochs on a single NVIDIA A800 GPU.

Object Detection in Darkness

COCO → ExDark Adaptation. We first evaluate our method on a challenging low-light object detection task: adapting from the well-lit COCO dataset to the low-light ExDark dataset. In this setting, COCO serves as the labeled source domain, while ExDark provides only unlabeled target domain images. No target annotations are used during training, which reflects a realistic deployment scenario where acquiring labeled data under low-light conditions is expensive or impractical.

Method	mAP@50:95	mAP@50
Source Only	32.2	63.4
Target Only	44.9	72.9
DLDA (Ours)	**36.8**	**68.8**

Table 1: Cross-domain detection results on ExDark test set (COCO → ExDark). DLDA significantly outperforms the source-only baseline and approaches the target-supervised upper bound, demonstrating its effectiveness in unsupervised domain adaptation.

Tab. 1 reports the detection performance on the ExDark test set. Our DLDA model achieves 36.8mAP@50:95 and 68.8mAP@50, significantly outperforming the source-only baseline (32.2 / 63.4), and substantially narrowing the gap to the target-supervised upper bound (44.9 / 72.9). These results demonstrate that our dual-level domain adaptation framework effectively mitigates the domain shift caused by extreme illumination degradation in a fully unsupervised setting.

Comparison with State-of-the-Art Methods. To further assess the effectiveness of DLDA, we compare it with several state-of-the-art methods under the same COCO → ExDark adaptation scenario, as shown in Tab. 2. We select representative methods from three categories: domain adaptation methods (MS-DAYOLO (Hnewa and Radha 2021), DAI (Du, Shi, and Deng 2024)), low-light enhancement methods (MBLLEN (Lv et al. 2018), KIND (Zhang et al.

Method	Bicycle	Boat	Bottle	Bus	Car	Cat	Chair	Cup	Dog	Motorbike	People	Table	Total
Unsupervised Domain Adaptation (without target labels)													
YOLO(source only)	69.1	39.1	62.3	86.6	76.3	68.4	58.8	50.4	76.1	64.5	76.9	32.6	63.4
MS-DAYOLO	73.9	45.2	65.9	87.1	77.9	79.2	61.0	50.4	73.8	67.8	79.2	35.8	66.4
DAI	73.7	45.5	66.7	87.4	79.5	74.1	61.3	50.7	79.5	68.1	78.2	36.1	66.7
DLDA(Ours)	**77.8**	**47.6**	**68.8**	**89.0**	**81.6**	74.2	**63.4**	**52.8**	**80.8**	**70.2**	**81.3**	**38.2**	**68.8**
Fully Supervised (with target labels)													
YOLO(+FT)	85.5	74.0	71.1	91.2	83.0	74.5	70.0	75.9	80.0	86.2	78.8	63.1	77.8
MBLLEN	82.6	74.1	69.6	90.8	85.8	73.6	70.7	78.4	77.6	88.1	75.5	61.7	77.4
KIND	81.2	74.8	70.4	91.0	83.4	74.7	69.5	76.2	78.0	85.7	81.3	62.1	77.8
Zero-DCE	88.0	71.6	72.8	92.1	79.7	76.8	74.1	74.6	80.9	86.6	83.9	62.8	78.6
MAET	88.6	74.0	74.7	91.0	85.6	77.5	74.5	75.8	88.2	86.1	81.6	59.5	79.8
IAT	87.3	75.5	73.0	90.2	85.2	79.4	75.3	75.8	90.5	84.4	82.9	61.0	80.2
DAI(+FT)	87.8	74.8	73.8	91.6	86.4	80.9	74.5	79.3	89.2	85.8	83.0	63.9	80.9
DLDA(+FT)(Ours)	**89.4**	**77.7**	**76.4**	**93.8**	**89.4**	**82.4**	**79.6**	**82.2**	**91.6**	**88.9**	**85.4**	**66.5**	**83.6**

Table 2: Per-category AP@50 results on the ExDark test set. The upper block shows direct cross-domain evaluation without access to target labels during training. The lower block reports results using additional ExDark supervision.

Adaptation Scenario	Method	mAP
WIDER FACE → DARK FACE	Source-only	23.8
	MS-DAYOLO	25.9
	DAI	26.8
	DLDA (Ours)	**27.5**

Table 3: Cross-domain detection results on the face detection task under severe illumination changes.

2019), Zero-DCE (Guo et al. 2020)), and low-light feature learning methods (MAET (Cui et al. 2021), IAT (Cui et al. 2022)). For a fair comparison, all methods are reimplemented based on our YOLOv5-L baseline. Fully supervised methods are fine-tuned on the target domain with official pretrained weights.

In the unsupervised setting, DLDA consistently outperforms prior domain adaptation approaches. In the fully supervised fine-tuning setting, where target labels are used for additional training, DLDA also delivers the highest performance (83.6 mAP@50), surpassing both enhancement-based and adaptation-based methods. These results highlight the advantages of integrating image-level and feature-level adaptation: enhancement-based methods often introduce artifacts that undermine semantic consistency, while feature-only alignment may fail to handle severe appearance shifts. In contrast, DLDA enables robust domain alignment by extracting domain-invariant features. Furthermore, the learned representations offer strong initialization for downstream fine-tuning under limited supervision.

Qualitative Analysis. Fig. 3 presents qualitative detection results on representative samples from the ExDark test set. Additional visualizations are provided in the supplementary material. Compared to enhancement-based, feature-based, and domain adaptation methods, our model generates more accurate bounding boxes and more confident predictions. These results confirm that our unified framework effectively improves object detection under challenging low-light and cross-domain conditions.

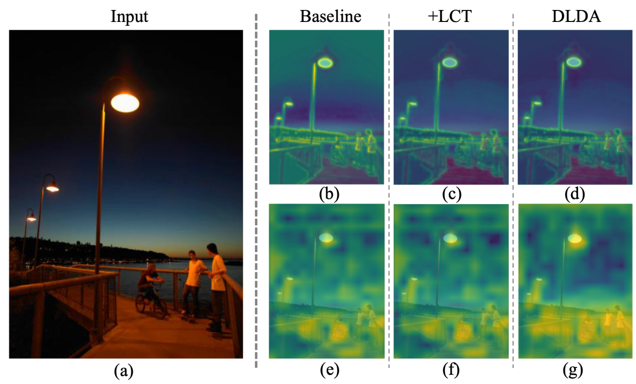


Figure 4: Feature visualization under different configurations. From left to right: input low-light image, features from Baseline, Baseline+LCT, and Full DLDA. Top row shows low-level features; bottom row shows high-level semantic features. Our DLDA model exhibits sharper, more concentrated responses, effectively suppressing background noise while focusing on foreground targets.

Generalization to Other Low-light Scenarios

To further evaluate the adaptability of the proposed DLDA framework, we conduct cross-domain experiments transferring from WIDER FACE to DARK FACE for face detection in extremely low-light conditions. This task is highly challenging due to the substantial degradation in image quality and the necessity for fine-grained detection. As shown in Tab. 3, DLDA consistently achieves the highest mAP, outperforming methods including MS-DAYOLO (Hnewa and Radha 2021) and DAI (Du, Shi, and Deng 2024). Notably, although DLDA is not specifically designed for face detection, it still demonstrates superior domain generalization and feature alignment capability, which is critical under severe illumination shifts. Qualitative comparisons in Fig. 5 further validate these findings. Compared to other methods, DLDA

(a) Baseline (b) MS-DAYOLO (c) DAI (d) Ours

Figure 5: Qualitative comparison on DARK FACE dataset (WIDER FACE → DARK FACE).

Method						mAP@50
Image-level			Feature-level			
DISP	CUT	LCT	$\mathcal{M}_1$	$\mathcal{M}_2$	MSCA	
–	–	–	–	–	–	63.4
✓	–	–	–	–	–	62.2
–	✓	–	–	–	–	65.5
–	–	✓	–	–	–	66.3
–	–	✓	✓	–	–	67.2
–	–	✓	–	✓	–	67.9
–	–	✓	–	–	✓	**68.8**

Table 4: Ablation results under the COCO→ExDark setting. Columns 1–3: image-level methods; columns 4–6: feature-level methods. ✓means the method is enabled. All results are mAP@50 on ExDark.

yields more accurate and comprehensive detection results in dark environments, with clearer localization and fewer missed faces, especially under occlusion or heavy shadows.

Ablation Study

To rigorously evaluate the effectiveness of each component in the proposed DLDA framework, we conduct ablation experiments under the COCO → ExDark setting. All models are trained without access to target domain annotations and evaluated on the ExDark test set using mAP@50 as the primary metric. The results are summarized in Tab. 4.

Image-level Adaptation. We first analyze the impact of image-level adaptation by comparing three low-light synthesis strategies: (1) DISP, a degradation-based synthesis method derived from MAET (Cui et al. 2021); (2) CUT (Park et al. 2020), a widely adopted contrastive translation approach; and (3) our proposed LCT module. When applied individually, DISP degrades performance below the source-only baseline, likely due to oversimplified degradations that disrupt semantic structures. CUT brings a moderate improvement (+2.1), while LCT achieves the highest gain (+2.9), demonstrating its effectiveness in generating illumination-adapted images.

Feature-level Adaptation. We then isolate the contribution of feature-level adaptation by evaluating three alignment strategies independently: multi-scale discriminators ($\mathcal{M}_1$ (Hnewa and Radha 2021)), conditional adversarial alignment ($\mathcal{M}_2$ (Long et al. 2018)), and our proposed MSCA. Each method yields improvements over the baseline, underscoring the importance of feature-space domain alignment. Among them, MSCA achieves the highest individual gain (+2.5), validating the effectiveness of multi-scale and class-conditional alignment in learning robust domain-invariant representations.

Effectiveness of Dual-level Adaptation. Combining LCT and MSCA achieves the best performance (68.8 mAP@50), significantly surpassing the source-only baseline (63.4) and all single-component variants. This confirms the complementary nature of the two adaptation levels: LCT mitigates appearance-level discrepancies, while MSCA facilitates semantic alignment in the feature space.

Feature Visualization. To gain insights into how different adaptation strategies affect feature extraction, we visualize intermediate feature maps from the backbone network across different settings. As shown in Fig. 4, we present the input low-light image alongside its corresponding feature responses under three configurations: (1) Baseline, (2) Baseline + Image-level Adaptation, and (3) Full DLDA with both LCT and MSCA.

The first row shows low-level features that capture fine-grained textures, while the second row presents high-level semantic features. In the Baseline model, the feature activations are weak and scattered due to the severe domain shift. With image-level adaptation, the network begins to highlight salient regions, but background interference remains. In contrast, our full DLDA model exhibits strong and concentrated activations around the foreground objects (e.g., the pedestrians), with clearer separation from the background and enhanced semantic consistency. These results demonstrate that dual-level adaptation not only enhances low-level structural perception but also improves high-level semantic focus, leading to more robust and discriminative feature representations under challenging low-light conditions.

Conclusion

In this paper, we propose DLDA, a unified dual-level domain adaptation framework for low-light object detection that jointly performs image-level style translation and feature-level semantic alignment. Luminance-aware Contrastive Translation (LCT) synthesizes target-style low-light images while preserving structural details, and Multi-scale Conditional Adversarial Alignment (MSCA) enforces class-conditioned alignment across multiple feature hierarchies to learn domain-invariant representations. Extensive experiments on COCO → ExDark and WIDER FACE → DARK FACE show that DLDA consistently surpasses state-of-the-art methods, and ablation studies verify the complementary benefits of LCT and MSCA. Notably, the MSCA branch is removed at inference, ensuring no additional computational burden. We believe DLDA provides an efficient and scalable solution for real-world low-light detection scenarios.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (Grant Nos. 62422314, 62273262 and 62088101), Shanghai Science and Technology Commission Project (Grant Nos. 24ZR1492700 and 2021SHZDZX0100), Aeronautical Science Foundation(2024M071038001), Industry-University-Research Cooperation Fund of the Eighth Research Institute of China Aerospace Science and Technology Corporation (SAST2023-019), China University Industry, University and Research Innovation Fund (Grant No. 2021ZYA03004), Eastern Talent Program, Shanghai Pilot Program for Basic Research, and Fundamental Research Funds for the Central Universities.

References

Arruda, V. F.; Paixao, T. M.; Berriel, R. F.; De Souza, A. F.; Badue, C.; Sebe, N.; and Oliveira-Santos, T. 2019. Cross-domain car detection using unsupervised image-to-image translation: From day to night. In *2019 International joint conference on neural networks (IJCNN)*, 1–8. IEEE.

Bhattacharjee, D.; Kim, S.; Vizier, G.; and Salzmann, M. 2020. Dunit: Detection-based unsupervised image-to-image translation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 4787–4796.

Carion, N.; Massa, F.; Synnaeve, G.; Usunier, N.; Kirillov, A.; and Zagoruyko, S. 2020. End-to-End Object Detection with Transformers. In *European Conference on Computer Vision*, 213–229.

Cui, Z.; Li, K.; Gu, L.; Su, S.; Gao, P.; Jiang, Z.; Qiao, Y.; and Harada, T. 2022. You Only Need 90K Parameters to Adapt Light: a Light Weight Transformer for Image Enhancement and Exposure Correction. In *Proceedings of the 33rd British Machine Vision Conference (BMVC)*, 238.

Cui, Z.; Qi, G.-J.; Gu, L.; You, S.; Zhang, Z.; and Harada, T. 2021. Multitask AET with Orthogonal Tangent Regularity for Dark Object Detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2553–2562.

Du, Z.; Shi, M.; and Deng, J. 2024. Boosting Object Detection with Zero-Shot Day-Night Domain Adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 12666–12676.

Everingham, M.; Van Gool, L.; Williams, C. K.; Winn, J.; and Zisserman, A. 2010. The Pascal Visual Object Classes (VOC) Challenge. *International Journal of Computer Vision (IJCV)*, 88(2): 303–338.

Fan, D.-P.; Cheng, M.-M.; Liu, Y.; Gao, S.-H.; Hou, Q.; and Borji, A. 2017. DARK FACE: A Face Detection Benchmark in Low-Light Environment. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 1–9.

Ganin, Y.; Ustinova, E.; Ajakan, H.; Germain, P.; Larochelle, H.; Laviolette, F.; Marchand, M.; and Lempitsky, V. 2016. Domain-adversarial training of neural networks. In *Journal of Machine Learning Research*, volume 17, 2096–2030.

Guo, C.; Li, C.; Guo, J.; Loy, C. C.; Hou, J.; Kwong, S.; and Cong, R. 2020. Zero-Reference Deep Curve Estimation for Low-Light Image Enhancement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 1777–1786.

Hashmi, K. A.; Kallempudi, G.; Stricker, D.; and Afzal, M. Z. 2023. Featenhancer: Enhancing hierarchical features for object detection and beyond under low-light vision. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 6725–6735.

He, K.; Gkioxari, G.; Dollár, P.; and Girshick, R. 2017. Mask R-CNN. In *Proceedings of the IEEE International Conference on Computer Vision*, 2961–2969.

Hnewa, M.; and Radha, H. 2021. MS-DAYOLO: Multi-scale domain adaptive YOLO for object detection. In *ICIP*, 338–342.

Hong, M.; Cheng, S.; Huang, H.; Fan, H.; and Liu, S. 2024. You Only Look Around: Learning Illumination-Invariant Feature for Low-light Object Detection. *Advances in Neural Information Processing Systems*, 37: 87136–87158.

Hong, M.; Li, S.; Yang, Y.; Zhu, F.; Zhao, Q.; and Lu, L. 2021. SSPNet: Scale selection pyramid network for tiny person detection from UAV images. *IEEE geoscience and remote sensing letters*, 19: 1–5.

Huang, S.-C.; Le, T.-H.; and Jaw, D.-W. 2020. DSNet: Joint semantic learning for object detection in inclement weather conditions. *IEEE transactions on pattern analysis and machine intelligence*, 43(8): 2623–2633.

Jocher, G.; et al. 2021. YOLOv5. https://github.com/ultralytics/yolov5. Accessed: 2025-08-03.

Li, F.; Xu, K.; Zhou, T.; and Zhang, S. 2022. DAYOLO: A Domain Adaptive YOLO for Robust Daytime and Nighttime Object Detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 3175–3184.

Li, G.; Wang, Y.; He, B.; Pang, T.; and Gao, M. 2025. Low-light multimodal object detection: A survey. *Computer Science Review*, 58: 100804.

Li, S.; Yang, Y.; Zeng, D.; and Wang, X. 2023. Adaptive and background-aware vision transformer for real-time uav tracking. In *Proceedings of the IEEE/CVF international conference on computer vision*, 13989–14000.

Liang, Z.; Li, C.; Zhou, S.; Feng, R.; and Loy, C. C. 2023. Iterative prompt learning for unsupervised backlit image enhancement. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 8094–8103.

Lin, C.-T.; Kew, J.-L.; Chan, C. S.; Lai, S.-H.; and Zach, C. 2023. Cycle-object consistency for image-to-image domain adaptation. *Pattern Recognition*, 138: 109416.

Lin, T.-Y.; Goyal, P.; Girshick, R.; He, K.; and Dollár, P. 2017. Focal Loss for Dense Object Detection. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2980–2988.

Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; and Zitnick, C. L. 2014. Microsoft COCO: Common Objects in Context. In *European Conference on Computer Vision (ECCV)*, 740–755. Springer.

Liu, W.; Ren, G.; Yu, R.; Guo, S.; Zhu, J.; and Zhang, L. 2022. Image-adaptive YOLO for object detection in adverse weather conditions. In *Proceedings of the AAAI conference on artificial intelligence*, volume 36, 1792–1800.

Loh, Y. P.; and Chan, C. S. 2019. Getting to know low-light images with the ExDark dataset. In *Computer Vision and Image Understanding*, volume 178, 30–42.

Long, M.; Cao, Y.; Wang, J.; and Jordan, M. I. 2018. Conditional adversarial domain adaptation. In *NeurIPS*, 1647–1657.

Luo, R.; Wang, W.; Yang, W.; and Liu, J. 2023. Similarity min-max: Zero-shot day-night domain adaptation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 8104–8114.

Lv, F.; Lu, F.; Wu, J.; and Lim, C. 2018. MBLLEN: Low-Light Image/Video Enhancement Using CNNs. In *Proceedings of the British Machine Vision Conference (BMVC)*, 220.

Ma, T.; Ma, L.; Fan, X.; Luo, Z.; and Liu, R. 2022. PIA: Parallel architecture with illumination allocator for joint enhancement and detection in low-light. In *Proceedings of the 30th ACM international conference on multimedia*, 2070–2078.

Park, T.; Efros, A. A.; Zhang, R.; and Zhu, J.-Y. 2020. Contrastive Learning for Unpaired Image-to-Image Translation. In *European Conference on Computer Vision (ECCV)*, 319–345. Springer.

Ren, S.; He, K.; Girshick, R.; and Sun, J. 2015. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. In *Advances in Neural Information Processing Systems*, 91–99.

Wang, W.; Yang, H.; Fu, J.; and Liu, J. 2024. Zero-reference low-light enhancement via physical quadruple priors. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 26057–26066.

Wu, Y.; Pan, C.; Wang, G.; Yang, Y.; Wei, J.; Li, C.; and Shen, H. T. 2023. Learning semantic-aware knowledge guidance for low-light image enhancement. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 1662–1671.

Yan, Q.; Feng, Y.; Zhang, C.; Wang, P.; Wu, P.; Dong, W.; Sun, J.; and Zhang, Y. 2024. You only need one color space: An efficient network for low-light image enhancement. *arXiv preprint arXiv:2402.05809*.

Yang, S.; Luo, P.; Loy, C. C.; and Tang, X. 2016. WIDER FACE: A Face Detection Benchmark. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 5525–5533.

Yang, W.; Yuan, Y.; Ren, W.; Liu, J.; Scheirer, W. J.; Wang, Z.; Zhang, T.; Zhong, Q.; Xie, D.; Pu, S.; et al. 2020. Advancing image understanding in poor visibility environments: A collective benchmark study. *IEEE Transactions on Image Processing*, 29: 5737–5752.

Yang, Y.; and Soatto, S. 2020. Fda: Fourier domain adaptation for semantic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 4085–4095.

Yang, Z.; Huang, J.; Chang, J.; Zhou, M.; Yu, H.; Zhang, J.; and Zhao, F. 2023. Visual recognition-driven image restoration for multiple degradation with intrinsic semantics recovery. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 14059–14070.

Yin, Y.; Xu, D.; Tan, C.; Liu, P.; Zhao, Y.; and Wei, Y. 2023. Cle diffusion: Controllable light enhancement diffusion model. In *Proceedings of the 31st ACM International Conference on Multimedia*, 8145–8156.

Zhang, B.; Suo, J.; and Dai, Q. 2023. A complementary dual-backbone transformer extracting and fusing weak cues for object detection in extremely dark videos. *Information Fusion*, 97: 101822.

Zhang, K.; Zhang, Z.; Li, Z.; and Qiao, Y. 2016. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE signal processing letters*, 23(10): 1499–1503.

Zhang, Y.; Zhang, J.; Guo, X.; Zhang, W.; and Ma, J. 2019. Kindling the darkness: A practical low-light image enhancer. In *Proceedings of the 27th ACM International Conference on Multimedia*, 1632–1640.

Zhang, Y.; Zhang, Y.; Liu, X.; Chen, Y.; Wu, Y.; Chu, X.; Ge, Z.; Wang, J.; Zhang, X.; and Shi, J. 2023. RT-DETR: Real-Time Detection Transformer. In *arXiv preprint arXiv:2304.08069*.

Zhao, Q.; Li, G.; He, B.; and Shen, R. 2025. Deep Learning for Low-Light Vision: A Comprehensive Survey. *IEEE Transactions on Neural Networks and Learning Systems*.