scientific reports

Check for updates

OPEN

NID-DETR: A novel model for accurate target detection in dark environments

Qingyuan Pan[1,3], Qiang Liu[1,3] & Wei Huang[1,2]✉

Target detection in low-light conditions poses significant challenges due to reduced contrast, increased noise, and color distortion, all of which adversely affect detection accuracy and robustness. Effective low-light target detection is crucial for reliable vision in critical applications such as surveillance, autonomous driving, and underwater exploration. Current mainstream algorithms face challenges in extracting meaningful features under low-light conditions, which significantly limits their effectiveness. Furthermore, existing vision Transformer models demonstrate high computational complexity, indicating a need for further optimization and enhancement. Initially, we enhance the dataset during model training to optimize machine vision perception. Subsequently, we design an inverted residual cascade structure module to effectively address the inefficiencies in the global attention window mechanism. Finally, in the target detection output layer, we adopt strategies to reduce concatenation operations and optimize small object detection heads to decrease the model parameter count and improve precision. The dataset is divided into training, testing, and validation sets in a 7:2:1 ratio. Validation on the low-light dataset demonstrates a reduction of 27% in model parameters, with improvements of 2.4%, 4.8%, and 2% in AP$_{50:95}$, AP$_{50}$, and AP$_{75}$, respectively. Our model outperforms both the best baseline and other state-of-the-art models. These experimental results underscore the effectiveness of our proposed approach.

Keywords Image enhancement, Target detection, Optimize machine vision perception

Target detection in low-light conditions is a critical challenge in computer vision, with broad applications in nighttime surveillance, security monitoring, and dynamic underwater environments. The inherent difficulties stem from insufficient illumination, low contrast, increased noise, and color distortion, all of which lead to significant loss of image detail. As a result, conventional target detection algorithms struggle to capture fine-grained features, adversely affecting detection accuracy and robustness[1,2]. Existing models and algorithms struggle to perform effectively in low-light conditions, primarily due to challenges associated with image enhancement and visual model optimization. While image enhancement techniques can improve the quality of low-light images to a certain extent, relying solely on these enhancements can introduce new complications. For instance, excessive enhancement may result in detail loss, artifact generation, or noise amplification, ultimately undermining the detection model's capacity to accurately recognize targets. Furthermore, the computational cost associated with vision Transformer models is prohibitively high, limiting their effectiveness primarily to standard scenarios. Therefore, to address the challenges of extracting meaningful features from dark images, a novel architecture must be developed specifically for low-light conditions.

To address these challenges, researchers have explored several approaches to solve the aforementioned problems, including image enhancement, algorithm performance optimization, and advancements in hardware to mitigate the effects of low-light conditions on target detection. However, hardware advancements often entail high costs and introduce uncertainties for future work due to their dependency and limited scalability[3,4].

While traditional target detection algorithms[5]have made significant progress to some extent, addressing the specific complexities of low-light scenes requires advancements in both computer vision models and image processing techniques. This paper focuses on addressing the unique challenges of weak light target detection and proposes enhancing detection performance under such conditions. The aim is to contribute to the growing knowledge base in the field of weak light target detection. Exploring innovative methods is crucial for improving the scientific understanding and practical applicability of target detection systems under challenging lighting conditions. This understanding essentially differentiates between human sensory vision and machine vision. The

[1]School of Computer Science and Engineering, Wuhan institute of Technology, Wuhan 430205, China. [2]Wuhan I-Boron Photoelectric Technology Co., Ltd, Wuhan 430205, China. [3]Hubei Provincial Key Laboratory of Intelligent Robots, Wuhan 430205, China. ✉email: huangw@wit.edu.cn

task of restoring low-light images is complex, requiring innovative algorithms focused on mitigating the adverse effects of insufficient illumination while preventing unnecessary light from affecting the original image. The fusion of image features has made certain contributions in this field[6,7], and essentially, this idea can be borrowed for the fusion of normal and low-light images.

This study focuses on improving image enhancement and optimizing Transformer models for practical applications, including segmentation. It explores the strengths of CNNs and Transformers and introduces a versatile low-light target detection algorithm[8]. We propose NID-DETR, which integrates Night-Enhance, an inverted residual cascade structure (iRMB-cascaded), and DetectHead to refine RTDETR[9] for real-time detection. Extensive experiments validate its effectiveness. Our key contributions are summarized in three main aspects.:

1. A three-layer decomposition network is employed for image enhancement, which incorporates a Laplacian filter, light smoothing techniques, and an unsupervised loss function. This approach effectively separates the light effect layer from the background layer, generating high-quality images while minimizing the loss of target region information during the enhancement and restoration processes. Consequently, it provides deep learning models with a more accurate and realistic representation of the target.
2. A novel CNN-Transformer hybrid architecture is proposed, wherein the network employs a cascaded mechanism to progressively compute feature information for each attention head. This design enhances feature representation sequentially across successive heads, thereby improving model efficiency while maintaining a balanced computational overhead.
3. In the model's output architecture, an efficient DetectHead output is designed to eliminate redundancy and reduce excessive feature concatenation. Furthermore, a bidirectional data flow mechanism is introduced, which incorporates global modeling capabilities into the global branch to enhance feature representation and detection accuracy.

Related work

The field of image enhancement lies at the intersection of human visual perception and machine vision, addressing inherent differences in perceiving and processing visual information. Human visual perception tends to favor scenes with ample natural light, while machine vision, especially in deep learning, faces challenges such as noise and local detail distortion associated with strong lighting conditions. Utilizing low-light enhancement and high-light suppression helps restore the most realistic colors and provides optimal inputs for deep learning models. Currently, mainstream image enhancement models include LCDNet[10], Zero-DCE[11], SNR[12] among others, comprising both supervised and unsupervised approaches. In the domain of target detection, the Transformer and YOLO stand out as exemplary models. Many researchers have achieved remarkable results using Transformer architecture, with Baidu's RTDETR outperforming all YOLO algorithms. Building upon this baseline, we propose a low-light target detection algorithm that combines CNN with the Transformer architecture. The network architecture diagram is illustrated in Fig. 1. We tailor the main backbone of the network with inverted residual group convolution and high-precision detection output to enhance model performance specifically for low-light target detection.

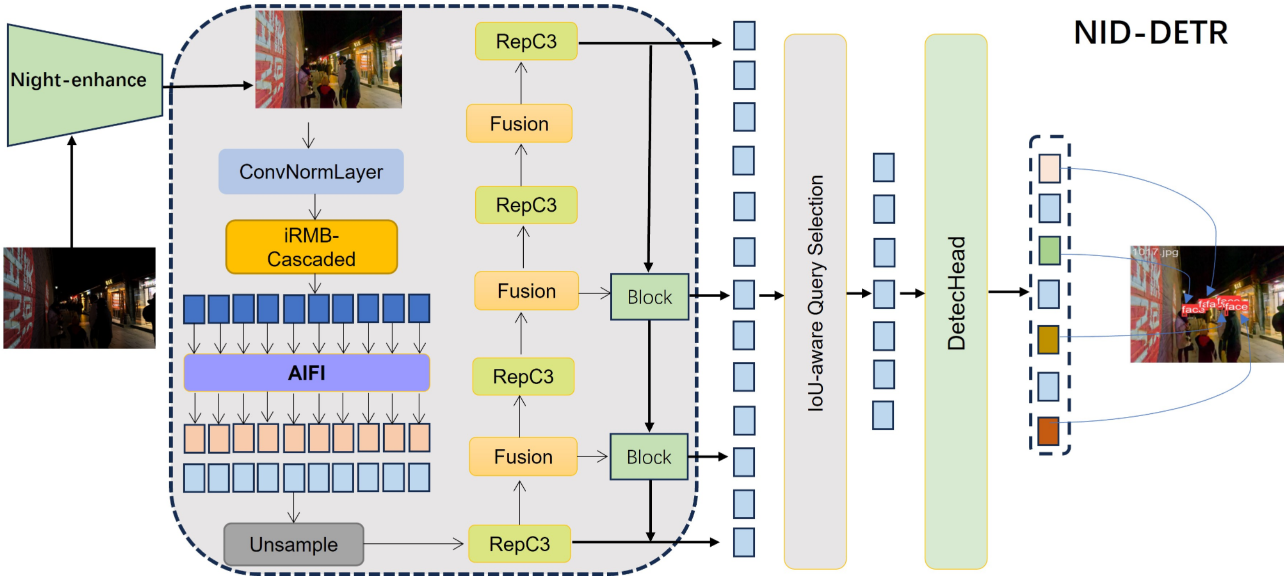


Fig. 1. Network architecture diagram of NID-DETR.

Method

Image enhancement network

Nighttime image enhancement is influenced not only by weak lighting but also by uneven distribution of strong light. The Night-enhance method consists of two layers: the layer decomposition network and the light suppression network[13]. The layer decomposition network utilizes unsupervised specific layer prior loss to separate shadow, reflection, and light effect layers. The light suppression network, guided by the estimated light effect layer, aims to suppress the light effect while enhancing the illumination of dark areas. To minimize artifacts and image detail distortion, structural and high-frequency consistency losses are introduced. In Formula (1), L , R, and G represent the shadow layer, reflection layer, and light effect layer respectively. I represents the input low-light image. The symbol of $\odot$ denotes element-wise multiplication of pixels in the image. Each of the three image layers is processed by its respective image layer network: light effect network $\phi\, G$, shadow network $\phi\, L$, and reflection network $\phi\, R$. They all utilize unsupervised loss to achieve a background scene with minimal light effect influence while enhancing contrast within the target area.

$$I = R \odot\ L + G \tag{1}$$

To improve clarity and flow without changing the meaning, we introduce two strategies for obtaining initial estimates of the light effect and shadow layers. Firstly, we calculate the maximum mapping $\mathrm{L_i}$ in the shadow layer part by selecting the maximum channel value among the three channels for each pixel. Then, smoothing techniques are applied to generate mapping $\mathrm{G_i}$. These steps lead to the formulation of the initial loss function formula (2) as depicted.

$$\mathcal{L}_{\text{init}}\ = |\mathrm{G} - \mathrm{G_i}|_1 + |\mathrm{L} - \mathrm{L_i}|_1 \tag{2}$$

During the image decomposition process, it is common for both the background layer image and the light effect layer image network to display conflicting distributions, with long-tailed and short-tailed states[14,15]. To address this inconsistency, use the integration of a gradient exclusion loss to refine unrelated image layers, with the main objective of separating these two layers. In Formula (3), a novel loss function is introduced, where $\mathrm{J_{init}}$ denotes the estimated background image obtained during decomposition, free from the influence of other light effects. Both $\mathrm{G}^{\downarrow\, n}$ and $\mathrm{J}_{\text{init}}^{\lfloor\, n}$ are derived through bilinear interpolation. To mitigate potential color distortion in the decomposition output, inspired by the gray world concept, a color constancy loss function is formulated in Formula (4), utilizing combinations of two channels in RGB denoted as c1 and c2.

$$\mathcal{L}_{\text{excl}}\ = \sum{}_{n=1}^{3} \|\ \tanh\left(\lambda_{\ \mathrm{G}\downarrow\, n} \left|\nabla\, \mathrm{G}^{\lfloor\, n}\right|\right) \circ \tanh\left(\lambda_{\ \mathrm{J}_{\text{init}}^{\lfloor\, n}} \left|\nabla\, \mathrm{J}_{\text{init}}^{\lfloor\, n}\right|\right) \|_{\ F} \tag{3}$$

$$\mathcal{L}_{\text{cc}} = \sum{}_{(c1,c2)} \left(\left|\mathrm{J}_{\text{init}}^{c1}\ - \mathrm{J}_{\text{init}}^{c2}\ \right|_1\right) \tag{4}$$

The essence of suppressing light effects involves utilizing data-driven methods, typically through training on paired data that includes both images with light effects and images without light effects. This training approach involves using binary classification and Class Activation Maps (CAM) to learn weights. These weights are then applied to multiply the feature maps and CAMs, producing attention feature maps. By analyzing these attention feature maps, the network's learning weights can be better understood, allowing them to focus on the light effect areas and effectively suppress intense light areas, ultimately producing high-quality images. In Eq. (5), the learning process is guided by identifying the presence of light effects. The generator and classifier of the network are denoted as $\varphi_{\ gen}$ and $\Gamma_{\ \text{gen}}$, respectively. The input to the network is J_{init}, which is connected to G. $\Gamma_{\ \text{gen}}$ acts as a classifier that categorizes features based on encoded features $f_{\text{e}} = (G, J_{init})$ and $f_{\text{ef}}\ = (G_0, J_{ef})$, with initial feature extraction carried out through average pooling. To enhance the suppression effect of light effects, an attention loss is introduced to guide the classifier $\Gamma_{\ \text{gen}}$ in focusing on feature maps with and without light effects. This design aims to produce specific feature maps that improve the suppression of light effects. A detailed schematic is shown in Fig. 2.

$$\mathcal{L}_{atten} = -\left(\mathbb{E}[\log(\Gamma_{\ \text{gen}}\ (f_{\text{e}})\ )] + \mathbb{E}[\log(1 - \Gamma_{\ \text{gen}}\ (f_{\text{ef}}\ )\ ))]\right) \tag{5}$$

Inverted residual cascade structure module

Large-scale Transformers-based visual detectors often come with significant system resource overhead. The primary advantage of iRMB-Cascaded lies in achieving an optimal performance-cost balance, enhancing the model's detection accuracy. This highly efficient hybrid model surpasses traditional Transformer models by enhancing dynamic modeling capabilities and reinforcing long-range interaction relationships[16,17]. While Multi-Head Self-Attention (MHSA) effectively captures various relationships within sequences through its multi-head mechanism, its significant computational overhead and resource consumption have driven current efforts towards reducing computational costs. In our network design, we introduce the iRMB-Cascaded network structure, as shown in Fig. 3. The redundancy issue in MHSA results in inefficient computation, which we address by incorporating a cascaded group attention mechanism into the model architecture. This mechanism provides different partitions of input features for each head, facilitating a clear decomposition of each head's attention for computation. The computation process can be represented by formulas (6) and (7). In these formulas, the jth head computes self-attention on X_{ij}, which represents the jth partition of input features. The projection layers

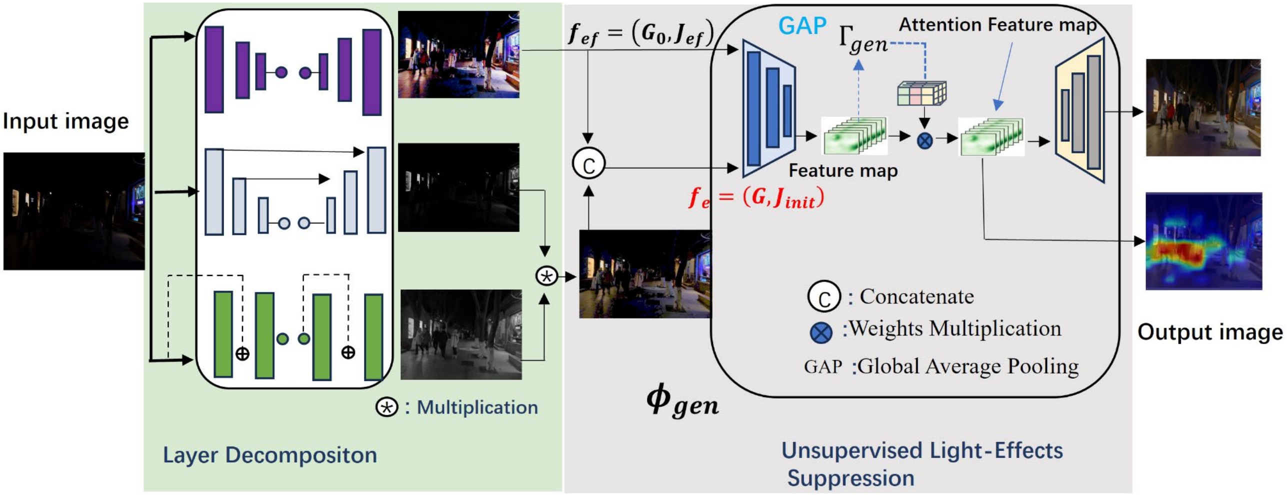
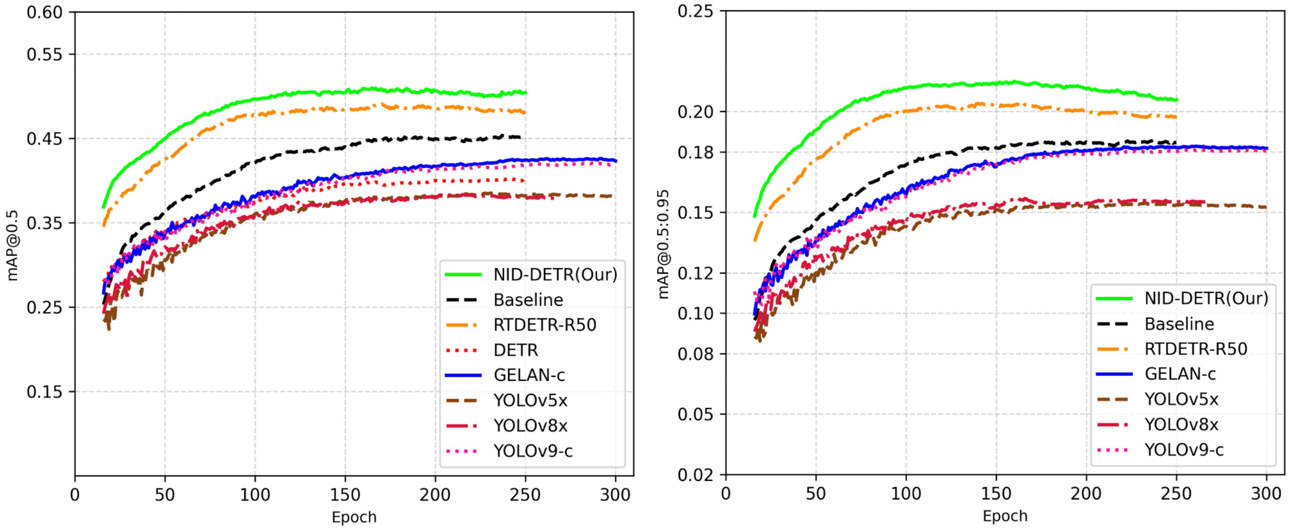


Fig. 2. The architecture of image enhancement network.



(a) mAP@0.5

(b) mAP@0.5:0.95

Fig. 3. mAP curve variations of mainstream model algorithms.

W_{ij}^{Q}, W_{ij}^{K}, and W_{ij}^{V} map input features to different subspaces and project the connected output features back to dimensions consistent with the input.

$$\widetilde{X}_{ij} = \mathrm{Attn}\left(X_{ij}W_{ij}^{\mathrm{Q}}, X_{ij}W_{ij}^{\mathrm{K}}, X_{ij}W_{ij}^{\mathrm{V}}\right) \tag{6}$$

$$\widetilde{X}_{i+1} = \mathrm{Concat}\left[\widetilde{X}_{ij}\right]_{j=1:h} W_i^{\mathrm{P}} \tag{7}$$

While utilizing feature segmentation in each attention head can reduce computational costs, it is essential to continuously improve its capacity by learning more complex feature projections at the Q, K, and V layer. Employing a cascaded approach to calculate the attention map for each head, gradually incorporating the output of each head into subsequent heads, refines the feature representation incrementally. The key idea behind this strategy is to introduce additional information to enhance the expressive ability of each attention head, ultimately enhancing the learning capacity of the entire module.

$$X'_{ij} = X_{ij} + \widetilde{X}_{i(j-1)}, 1 < j \leq \ h \tag{8}$$

In Formula (8), X'_{ij} represents the combination of the jth input segment X_{ij} and the output of the previous head. After the Q projection, a token interaction layer is introduced to enable the self-attention mechanism to capture both local and global relationships simultaneously. This cascaded design provides two main benefits: firstly, each attention head focuses on different feature segments, increasing the diversity of attention maps. Secondly, cascaded attention reduces the number of channels in the input and output of the Q, K, and V layer, effectively decreasing the parameter count. As illustrated in the Table 1, the model's parameters are reduced by 27% compared to the Baseline. Another advantage is that by cascading attention heads, the network's depth is increased, thereby enhancing the model's capacity. Subsequent to the cascaded outputs, dilated convolutions are incorporated into the network model structure to leverage an expanded receptive field for handling contextual relationship tasks. Furthermore, SE attention[18]is separately applied to the output feature layers of reshape and Dilation-Conv[19]. This attention mechanism gathers global information through global pooling and utilizes fully connected layers to generate channel attention weights, thereby enhancing the network's sensitivity to key features and improving model performance. The detailed network architecture is shown in Fig. 4.

Efficient output prediction layer with reduced redundancy

In the RTDETR model, the fusion of encoded information from the Decoder and Head parts occurs after the concat module and in the CCFM module[9]. This fusion involves combining features from top-down and bottom-up approaches through concatenation. While this method allows for the integration of multiple features, it also increases the dimensionality of each channel due to the concatenation of feature maps. The simultaneous processing of information from these maps can lead to higher computational demands, especially in large-scale models, posing challenges in scenarios with limited computational resources. This can result in increased computational costs during both training and inference. Uneven distribution or inconsistency of feature information in space or channels may further impact the model, particularly in ambiguous situations like low-light conditions where the distinction between foreground and background is unclear. To address these issues, our improvement strategy focuses on reducing half of the feature fusion and generating additional prediction outputs[20]. We have incorporated Repc3 and Block modules as the output features in the network structure, resulting in a total of four modules. The Block module comprises conventional convolution and Repc3 modules, aiming to strike a balance between model expressiveness, computational efficiency, and parameter reduction. This approach ensures optimal performance and scalability in complex scenarios. The network architecture part can refer to Fig. 1.

Experimental result

Dataset and experimental environment configuration

In our study, we employed the publicly available DARK FACE[21] dataset to validate our proposed approach. This dataset consists of 7000 images taken in low-light and dark conditions. The dataset was split into training, testing, and validation sets with 4900, 1400, and 700 images, respectively, following a ratio of 7:2:1. Evaluation metrics for the models included floating-point operations, parameter count, mAP at IoU thresholds of 0.5 and 0.95, as well as COCO AP validation metrics. Our experiments were conducted on the Ubuntu 20.04 operating

Model	FLOPS/G	Params/M	AP$_{50:95}$	AP$_{50}$	AP$_{75}$	AP$_{S}$	AP$_{M}$	AP$_{L}$
YOLOv5 m	64.0	25.04	14.4%	35.2%	9.6%	12.5%	37.2%	38.5%
YOLOv5 l	134.7	53.1	14.8%	36.1%	9.9%	12.9%	37.4%	39.2%
YOLOv5x	246.0	97.2	15%	36.8%	9.9%	13.1%	37.4%	40.7%
YOLOx	27.0	9.0	13.1%	33.9%	7.6%	11.2%	35.0%	37.1%
YOLOv7	105.2	37.2	14.6%	34.9%	8.1%	12.0%	36.2%	36.7%
DETR (ResNet50)	86.0	41.0	15.3%	37.4%	9.9%	14.3%	48.5%	48.9%
YOLOv8 m	78.7	25.8	14.6%	36.2%	9.7%	12.9%	36.2%	38.8%
YOLOv8 l	164.8	43.6	15.0%	36.5%	9.8%	13.1%	36.6%	40.8%
YOLOv8x	257.4	68.1	15.0%	36.4%	10.2%	13.0%	38.0%	40.2%
YOLOv9	18.3	4.2	14.3%	37.5%	9.8%	13.0%	37.6%	38.1%
YOLOv9-c	237.6	50.9	16.0%	38.6%	10.6%	14.8%	39.9%	39.8%
YOLOv9-e	243.3	69.3	17.2%	39.6%	12.0%	16.1%	41.2%	41.6%
GELAN-c	102.5	25.4	16.2%	37.9%	10.3%	14.7%	40.1%	40.6%
GELAN-e	190.8	58.0	18.6%	39.8%	11.9%	16.1%	42.0%	42.1%
RTDETR(ResNet18)	56.9	19.8	18.9%	45.2%	13.1%	17.5%	35.9%	35.9%
RTDETR(ResNet34)	87.1	30.0	19.6%	46.9%	15.0%	18.7%	36.5%	37.1%
RTDETR(ResNet50)	129.5	42.0	20.3%	47.8%	14.2%	19.3%	52.2%	46.9%
RTDETR(ResNet101)	247.1	74.7	19.7%	46.9%	13.4%	18.6%	50.8%	47.6%
DINO[39]	279	47	19.5%	47.8%	14.2%	19.1%	49.3%	41.9
FeatEnHancer[17]	-----	-----	19.9%	47.2%	-----	-----	-----	-----
NID-DETR(Our)	68.0	14.0	21.3%	50.0%	15.1%	20.4%	50.7%	42%

Table 1. Different algorithm models compare experimental results.

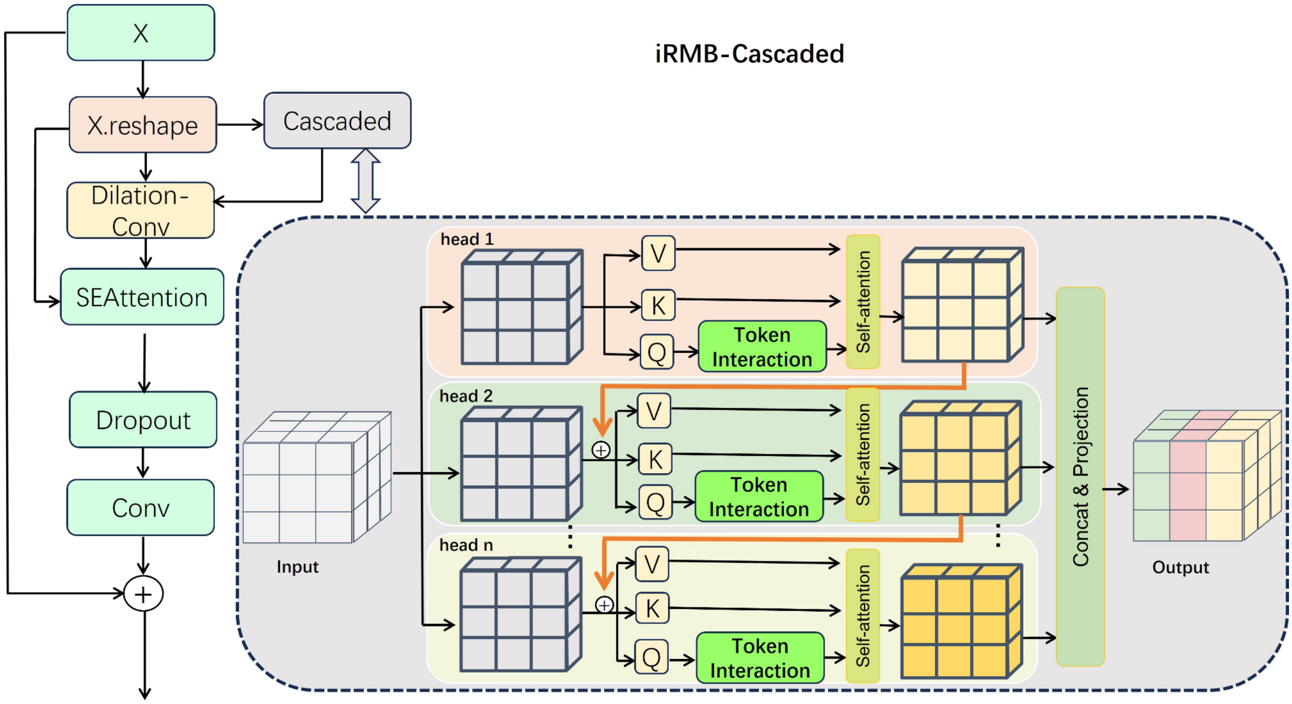


Fig. 4. Structure diagram display of Inverted residual cascade structure module.

Method	$AP_{50:95}$	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
Baseline	18.9%	45.2%	13.1%	17.5%	35.9%	35.9%
LCDNet	19.1%	46.6%	12.6%	17.7%	36.5%	39.4%
Zero-DCE	19.4%	46.9%	13.3%	18.1%	37.2%	39.7%
SNR	19.3%	46.6%	12.9%	17.7%	38.3%	41.6%
Night-enhance	19.5%	46.6%	13.3%	18.6%	49.4%	50.7%

Table 2. Comparative results from image enhancement ablation experiments.

system and utilized the PyTorch deep learning framework. The experiments were carried out using four RTX 2080 ti GPUs.

Comparative experiments

The comparison of mainstream detectors on the low-light target detection dataset was conducted, with the Baseline being RTDETR[9]using ResNet18 as the backbone. Alongside our NID-DETR, we evaluated other excellent object detection models such as YOLOv5, YOLOx[22], YOLOv7[23], DETR[24], YOLOv8, YOLOv9[25], GELAN, DINO[38] and RTDETR. Detailed experimental results are provided in Table 2. RTDETR, based on ResNet50, achieved relatively higher AP metrics, with AP_M and AP_L being 1.5% and 4.9% higher than our NID-DETR, respectively. However, its model parameters and floating-point operations are approximately double and four times that of our model. In real-time object detection scenarios, such overheads are impractical, presenting significant challenges for deployment in practical applications. In terms of model parameters and computational cost, our model demonstrates superiority. Our model outperforms DETR (ResNet50) by 6%, 12.6%, and 5.2% in $AP_{50:95}$, AP_{50}, and AP_{75}, respectively. Compared to YOLOv8x, YOLOv9e, and GELAN-e, our model surpasses them by 6.3%, 13.6%, and 4.9%; 4.1%, 10.4%, and 3.1%; and 2.7%, 10.2%, and 3.2%, respectively. The model's parameter counts and floating-point operations are approximately one-fourth and one-fifth of these state-of-the-art (SOTA) models, respectively. Extensive data analysis suggests that even the powerful CNN-based YOLOv9 and GELAN models cannot comprehensively outperform Transformer-based object detectors in complex low-light conditions. We also compared our model with the Featenhancer[26] model (ICCV2023) and maintained a leading advantage in terms of $AP_{50:95}$ and AP_{50}. NID-DETR, based on a self-attention mechanism, effectively captures global information of input sequences, allowing each position to attend to information from other positions, thereby better handling global dependencies. It demonstrates strong adaptive capabilities across various domains and tasks, even in complex low-light environments. Beyond low-light object detection, our two proposed strategies offer significant advantages in real-world applications compared to other methods. While the NID-DETR model may exhibit slightly lower detection accuracy for large objects compared to the baseline model, our analysis suggests that this limitation is closely related to the image enhancement process and the characteristics of the specific dataset used.

Analysis of mAP variation in mainstream models

Research efforts in low-light target detection include Yin et al.[27]. integrated the Laplacian Enhancement Pyramid Network with YOLOv3[28]for the ExDark[29] dataset, achieving high detection accuracy and speed. Wang et al.[30]. addressed the issue of low efficiency in image labeling under low-light conditions by proposing the HLA-FACE network detection framework, utilizing bidirectional low-level adaptation and multitask high-level adaptation methods to resolve semantic information disparities across different levels. Cui et al.[31]. explored low-light detection tasks from the perspectives of physical noise modeling and image signal processing, designing the MAET model structure for low-light target detection using encoding and decoding mechanisms. Hashmi et al.[26]. proposed the FeatenHancer model, featuring novel and plug-and-play characteristics, which leverages task-specific loss functions to guide hierarchical combination of multi-scale features through multi-head attention mechanisms, achieving excellent performance across various low-light datasets. Based on their experiences, we design the general-purpose NID-DETR model for low-light target detection. The overall model structure is depicted in Fig. 1. Compared with mainstream detectors such as DETR, RTDETR (Baseline), YOLOv5, YOLOv8, YOLOv9, and GELAN, our NID-DETR model achieves the highest mAP with relatively fewer model parameters and floating-point operations, as detailed in Table 2. The mAP curve changes are illustrated in Fig. 3.

Ablation study

Restoring authentic images poses a significant challenge, as image enhancement techniques have the potential to improve image quality. However, excessive enhancement can negatively impact algorithms, affecting both CNN-based feature extraction and Transformer-based position encoding processes. To assess the effectiveness of various enhancement methods, including the original dataset, we conducted comparative analyses of their enhancement effects. Different models displayed distinct enhancement results, as shown in Fig. 5. While the original low-light image in Fig. 5(a) highlights inherent challenges, Fig. 5(d) demonstrates the superior visual enhancement achieved by SNR, a prominent model in the field, albeit with some instances of excessive local exposure. On the other hand, Fig. 5(c) struggles to address image enhancement in complex scenes, with some images showing less noticeable improvements. Additionally, Fig. 5(b) exhibits partial loss of details and distortion in the car image, which is located in the lower right corner of the image. Noteworthy is the more balanced enhancement effect in Fig. 5(e). Evaluation based on Table 3 indicates that Night-enhance outperforms the Baseline, enhancing $AP_{50:95}$, AP_{50}, and AP_{75} by 0.6%, 1.4%, and 0.2%, respectively. Furthermore, Night-enhance surpasses LCDNet, Zero-DCE, and SNR in terms of AP_S, AP_M, and AP_L metrics, with improvements of 0.9%, 12.9%, and 11.3%, 0.5%, 12.2%, and 11%, as well as 0.9%, 11.1%, and 9.1%, respectively. Night-enhance excels in refining low-light scenes and suppressing highlights, thereby providing optimal visual effects for deep learning models.

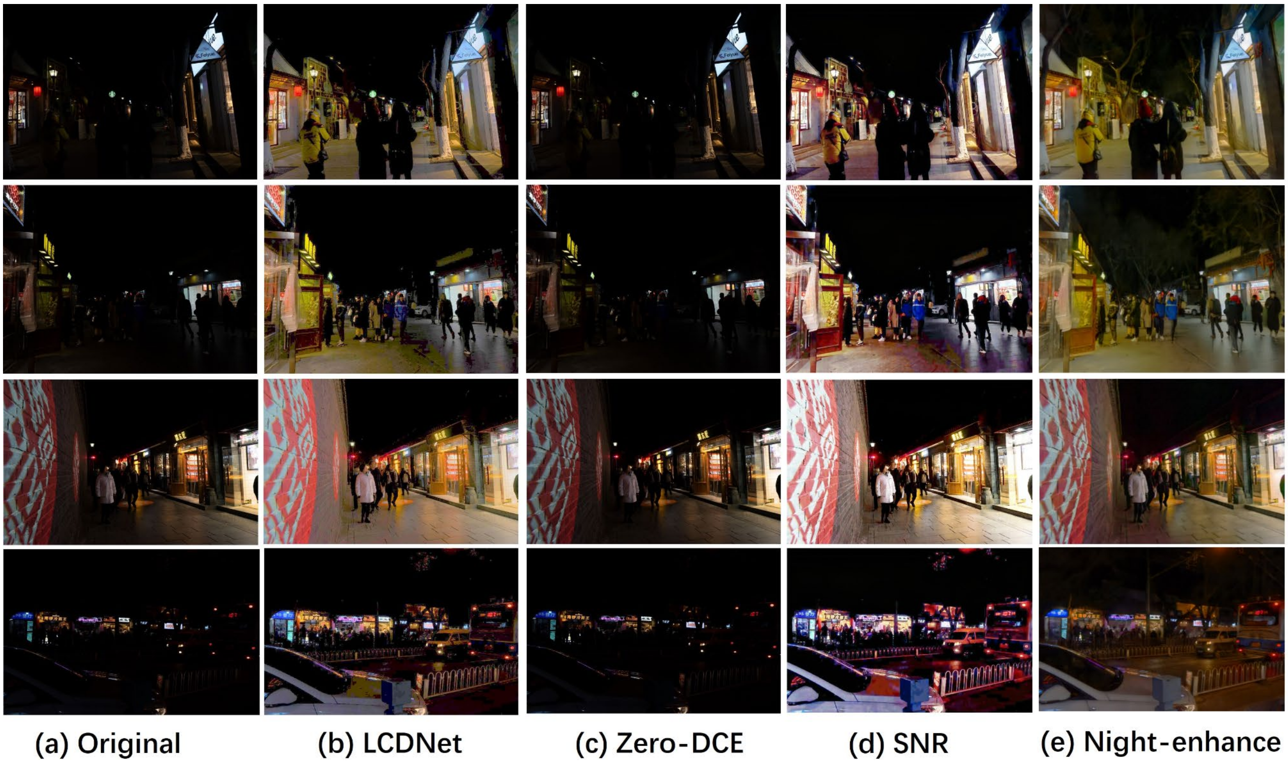


Fig. 5. Illustration of different enhancement effects.

Night-enhance	iRMB-Cascaded	DetectHead	Flops/G	Params/M	$AP_{50:95}$	AP_{50}	AP_{75}
Baseline			56.9	19.8	18.9%	45.2%	13.1%
√			56.9	19.8	19.5%	46.6%	13.3%
	√		57.0	17.7	19.1%	45.7%	13.2%
		√	59.3	18.7	19.5%	47.1%	13.2%
√	√		49.1	16.4	19.5%	47.5%	12.6%
√		√	78.1	18.6	21.1%	49.6%	15.1%
	√	√	46.8	15.2	19.7%	47.4%	13.0%
√	√	√	68.0	14.0	21.3%	50.0%	15.1%

Table 3. Comparative analysis of ablation experiment results for different image enhancements.

Method ablation experiments

To validate the efficacy of our approach, we provide comprehensive experimental data for each method utilized in Table 1. Although a minor imperfection was noted in the ablation experiments, our approach shows a slight increase in FLOPS metrics compared to both the baseline and other individual methods, indicating an increased demand for computational resources. This imperfect aspect is also the focus of our future work. Other methods, such as Night-enhance, iRMB-cascaded, and DetectHead, exhibit significant enhancements with improvements in $AP_{50:95}$, AP_{50}, and AP_{75} relative to the baseline. The results of the ablation experiment strongly support the effectiveness of our method. Particularly, the use of image enhancement techniques leads to higher overall AP values compared to scenarios without such methods. Moreover, iRMB-Cascaded proves effective in enhancing model performance by drawing inspiration from group convolution design principles in CNNs, resulting in a reduction of 5.8 M parameters compared to the Baseline. Figure 6 displays the fluctuations in mAP and loss curves for each method in the ablation experiments. Our proposed methodology not only achieves the lowest loss during training but also achieves higher mAP values after model convergence.

Comparison experiment of mainstream models

The contemporary evolution of Transformer model architectures reveals a diverse and noteworthy trajectory, marked by the emergence of several exemplary contenders, among which stand numerous robust backbone networks. Notable among these is the SwinTransformer[32]proposed by the Microsoft team, characterized by its hierarchical visual model featuring a flexible window mechanism. Additionally, the Fasternet[33]architecture improved accuracy and reduced FLOPS requirements. ConvNexts[34]introduces groundbreaking advancements in convolutional masking and encoder frameworks, revolutionizing the landscape of model architectures. While the Efficientformery[35] model garners widespread attention owing to its innovative fine-grained joint search strategy. These models, which integrate Convolutional Neural Networks (CNNs) with Transformers, demonstrate significant potential. We utilize them as baseline models for the RTDETR backbone and assess their performance using the mAP_{50} and $mAP_{50:95}$ metrics. As illustrated in Table 4, RTDETR with FasterNet and EfficientFormer backbones achieves commendable mAP scores while simultaneously reducing FLOPS and parameters. Notably, NID-DETR surpasses these models by approximately 4.7% and 4.1% in mAP_{50}, and by 2.6% and 2.5% in $mAP_{50:95}$, respectively. Although NID-DETR has an additional 1 M to 2 M parameters, it continues to provide outstanding mAP performance.

Our study presents a thorough comparison with the latest models in the field, clearly demonstrating the efficacy of our approach, as delineated by the data provided in Table 5. We also referenced the HLA-FACE[30] model introduced by Wang et al. at CVPR 2021, acknowledging their notable contributions to the domain. Upon comparison, the table summarizes the advancements in low-light object detection in recent years, providing a comparative analysis of both fully supervised and semi-supervised methods. NID-DETR achieves the highest mean Average Precision (mAP) value of 50.9%, outperforming other methods. This result highlights the effectiveness of NID-DETR in low-light conditions. Compared to enhancement-based and darkening-based approaches, NID-DETR demonstrates superior robustness and adaptability, likely due to its specialized design for handling complex illumination variations. This suggests that our model successfully integrates enhancement and detection, leading to improved feature extraction and overall performance in challenging lighting environments.

Comparative analysis of heat maps

Visualizing the model's attention through heat maps provides an intuitive means to analyze its focus and decision-making processes. In Fig. 7, we compare the GradCAM visualizations of the original DARKFACE dataset with its enhanced counterpart. The results indicate that under low-light conditions, the model exhibits weak and dispersed activations, attributed to the lack of informative features and low contrast, which complicates object recognition. Following image enhancement, the model's attention becomes more concentrated, particularly on key objects such as pedestrians and illuminated areas. This underscores the effectiveness of image enhancement in improving feature extraction, thereby enhancing both model interpretability and object detection performance in challenging low-light environments.

Analysis of feature map visualization effects

From both theoretical analysis and experimental data performance, our work may appear somewhat abstract. To present our contributions in a more detailed and concrete manner, we chose to conduct an in-depth analysis

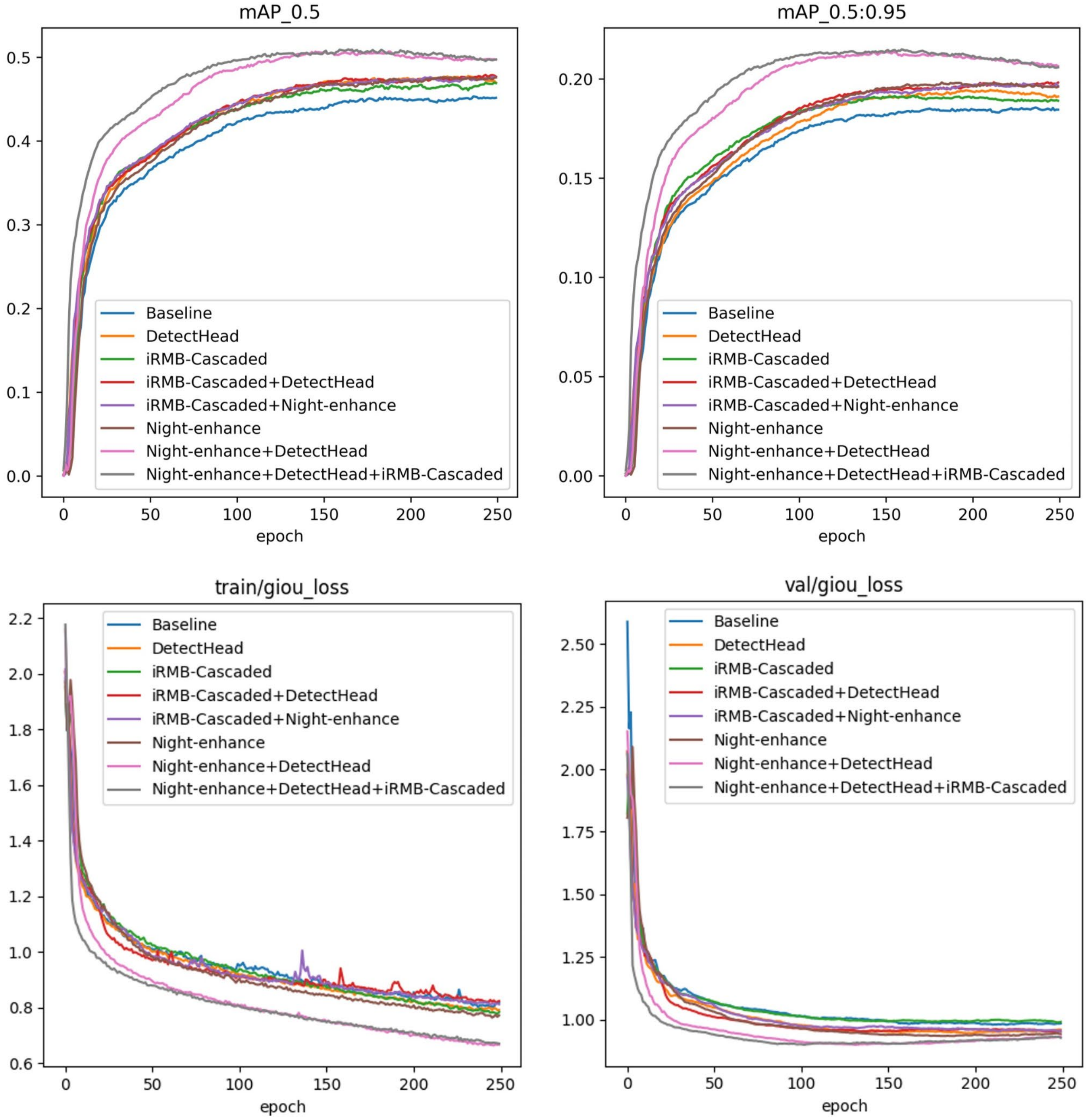


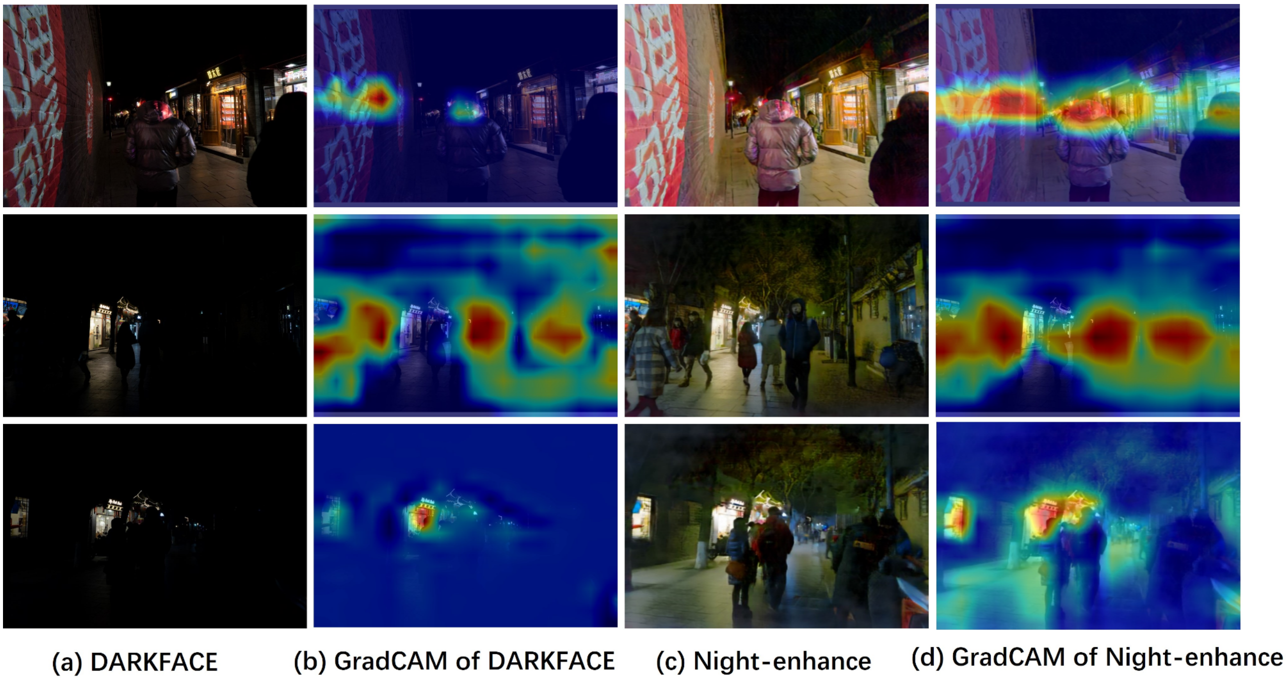
Fig. 6. Presentation of curves depicting mAP and loss calculation in ablation experiments.

Model	Backbone	Publish	Flops/G	Params/M	mAP@0.5	mAP@0.5:0.95
RT-DETR	SwinTransformer	ICCV2021	97.0	36.3	43.9%	17.4%
	FasterNet	CVPR2023	29.6	11.0	46.2%	18.9%
	ConvNextv2	CVPR2023	31.9	12.3	42.1%	17.2%
	EfficientformeryV2	ICCV2023	30.5	12.0	46.8%	19.0%
NID-DETR	/	/	68.0	14.0	50.9%	21.5%

Table 4. Comparative experiments on mainstream transformer models used as backbones (To ensure experimental accuracy, the datasets in Table 4 are all enhanced using the Night-enhance method for comparison).

Category	Method	mAP
Enhancement (with Small Hard Face)	Zero-DCE	37.7%
	MF[37]	38.3%
Enhancement (with DSFD)	LIME[38]	40.7
	Zero-DCE	41.3%
	MF	41.4%
Darkening (with DSFD)	MUNIT	29.7%
	CycleGAN[40]	31.9%
	CUT[41]	32.7%
Fully Supervised	Fine-tuned DSFD[42]	46.0%
-------------------------	HLA-FACE(CVPR2021)	44.4%
-------------------------	NID-DETR(Our)	50.9%

Table 5. Comparison results of different methods on the DARK FACE dataset (The experimental results in this table are referenced from the HLA-FACE, which does not provide detailed explanations for mAP. For consistency and rigor, only mAP is indicated in the table).



(a) DARKFACE (b) GradCAM of DARKFACE (c) Night-enhance (d) GradCAM of Night-enhance

Fig. 7. Visualization of heat maps comparison and analysis.

starting from the feature maps during the model training process. Towards the end of our study, we reproduced the PENet network proposed by Yin et al.[27], which includes detailed processing modules and low-frequency enhancement filters to improve various low-light images by constructing a Laplacian pyramid enhancement network. This network typically utilizes traditional enhancement methods to emphasize the distinction between foreground objects and the background, but it overlooks the potential impact of lighting and color on the model. The feature map display of this network is illustrated in Fig. 8(a). To gain deeper insights into the feature extraction process of the model, we compared the feature maps during the training process of the YOLOv8 and YOLOv9 models, as depicted in Figs. 8(b) and 8(c), respectively. Figure 8(d) highlights NID-DETR's clear feature maps, precise object contours, and strong target-background separation, enabling robust feature extraction and superior object detection performance, especially in low-light or complex scenarios. Upon observation, it is evident that as the training progresses, the details in the feature maps gradually become blurred, and the specific information starts to fade. In contrast, NID-DETR shows relatively intact feature map representation, with detailed information clearly preserved. This suggests that in terms of feature extraction and loss computation, NID-DETR offers more reliable identification information compared to other models[42].

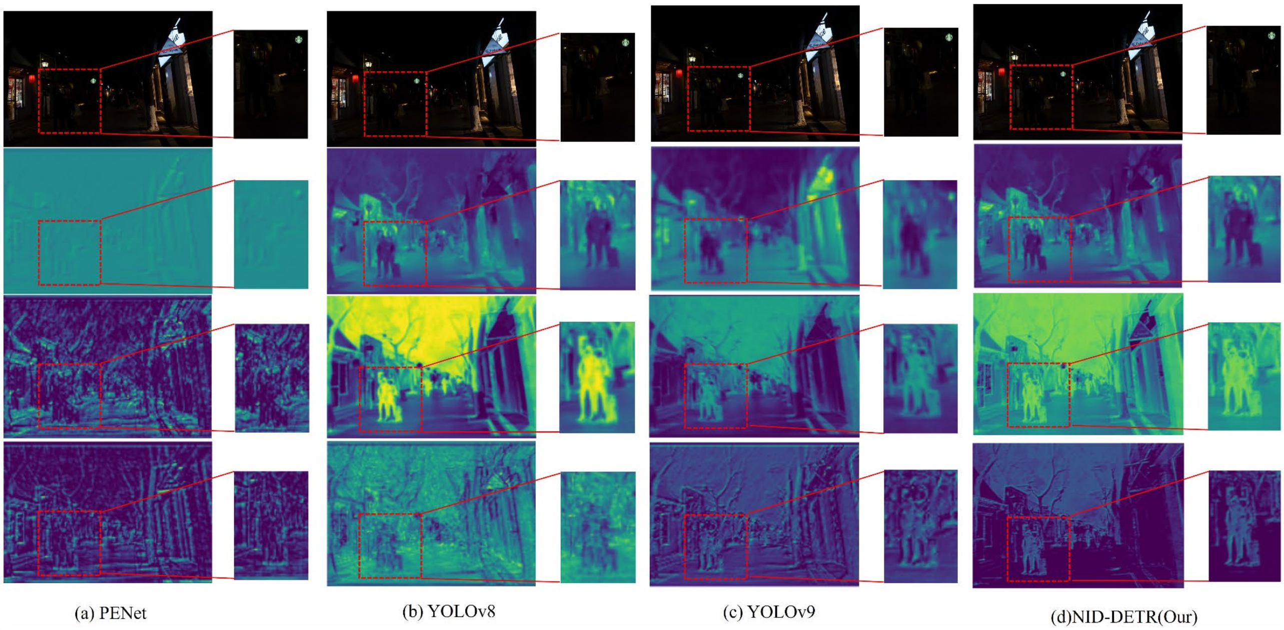


Fig. 8. Comparison of visualization results of feature maps from different models.

Conclusion

In our study, we present the NID-DETR model, which combines CNN and Transformer architecture to tackle challenges in low-light conditions. Our approach includes the development of an efficient inverted residual cascade structure and careful design of the network's prediction output layer. Results on the DARK FACE dataset show significant improvements in AP and mAP metrics compared to Baseline models and other state-of-the-art models. Additionally, our method achieves a notable 27% reduction in model parameters. When compared to peer achievements, NID-DETR remains at the forefront. Furthermore, through parameter optimization, it demonstrates a more favorable utilization of computational resources in practical scenarios, exhibiting strong competitiveness in diverse and complex conditions. Our model performs well in consistent low-light conditions but may struggle in highly dynamic lighting, such as transitions from indoor to outdoor nighttime scenes. Variations in light distribution and enhancement consistency affect its robustness. Ensuring adaptability across diverse environments remains an important direction for future research.

Future work

Our study emphasizes the significance of image enhancement for object detection and recognition in low-light settings. Furthermore, our experimental results highlight the difference between human visual perception and machine vision, pointing out the negative impact of excessive lighting. This observation stresses the need to adjust machine vision systems to function effectively in environments that replicate human perceptual conditions. Nevertheless, our research also identifies areas that offer potential for further investigation and enhancement. While our model effectively reduces the number of parameters, there is a slight increase in floating-point operations. Addressing the computational overhead associated with floating-point operations presents an exciting avenue for future research. Tackling this challenge will facilitate more efficient utilization of computational resources, improving the practical application of our approach in diverse and complex scenarios

Data availability

The datasets used and/or analysed during the current study are available from the corresponding author on reasonable request.

Received: 30 December 2024; Accepted: 9 April 2025
Published online: 08 May 2025

References

1. Xu, X., Wang, S., Wang, Z., Zhang, X. & Hu, R. Exploring image enhancement for salient object detection in low light images. *ACM Trans. Multimedia Comput. Commun. Appl. (TOMM)*. **17**, 1–19. https://doi.org/10.1145/3414839 (2021).
2. Wei, C., Wang, W., Yang, W. & Liu, J. Deep retinex decomposition for low-light enhancement. *ArXiv Preprint arXiv:1808 04560*. https://doi.org/10.48550/arXiv.1808.04560 (2018).
3. Hao, S., Wang, Z. & Sun, F. LEDet: A single-shot real-time object detector based on low-light image enhancement. *Comput. J.* **64**, 1028–1038. https://doi.org/10.1093/comjnl/bxab055 (2021).
4. Wu, Y. et al. Learning Semantic-Aware Knowledge Guidance for Low-Light Image Enhancement, Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition pp. 1662–1671. (2023).

5. Liu, W. et al. Ssd: Single shot multibox detector, Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14, Springer2016, pp. 21–37.
6. Singh, A. et al. Low-light image enhancement for UAVs with multi-feature fusion deep neural networks[J]. *IEEE Geosci. Remote Sens. Lett.* **19**, 1–5. https://doi.org/10.1109/LGRS.2022.3181106 (2022).
7. Wang, W. et al. An experiment-based review of low-light image enhancement methods[J]. *Ieee Access.* **8**, 87884–87917. https://doi.org/10.1109/ACCESS.2020.2992749 (2020).
8. Umirzakova, S. et al. Simplified knowledge distillation for deep neural networks bridging the performance gap with a novel Teacher–Student Architecture[J]. *Electronics* **13** (22), 4530 (2024).
9. Lv, W. et al. Detrs beat Yolos on real-time object detection. *ArXiv Preprint arXiv:2304 08069*. https://doi.org/10.48550/arXiv.2304.08069 (2023).
10. Wang, H., Xu, K. & Lau, R. W. Local color distributions prior for image enhancement, European Conference on Computer Vision, Springer2022, pp. 343–359.
11. Guo, C. et al. Zero-reference deep curve estimation for low-light image enhancement, Proceedings of the IEEE/CVF conference on computer vision and pattern recognition pp. 1780–1789. (2020).
12. Xu, X., Wang, R., Fu, C. W. & Jia, J. SNR-aware low-light image enhancement, Proceedings of the IEEE/CVF conference on computer vision and pattern recognition pp. 17714–17724. (2022).
13. Jin, Y., Yang, W. & Tan, R. T. Unsupervised night image enhancement: When layer decomposition meets light-effects suppression, European Conference on Computer Vision, Springer2022, pp. 404–421.
14. Li, Y. & Brown, M. S. Single image layer separation using relative smoothness, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition pp. 2752–2759. (2014).
15. Li, Y., Tan, R. T. & Brown, M. S. Nighttime haze removal with glow and multiple light colors, Proceedings of the IEEE international conference on computer vision pp. 226–234. (2015).
16. Zhang, J. et al. Rethinking mobile block for efficient attention-based models. *IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*. **IEEE Computer Society2023**, 1389–1400. https://doi.org/10.1109/ICCV51070.2023.00134 (2023).
17. Liu, X. et al. Efficientvit: Memory efficient vision transformer with cascaded group attention, Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition pp. 14420–14430. (2023).
18. Hu, J., Shen, L. & Sun, G. Squeeze-and-excitation networks, Proceedings of the IEEE conference on computer vision and pattern recognition pp. 7132–7141. (2018).
19. Wang, P. et al. Understanding convolution for semantic segmentation, 2018 IEEE winter conference on applications of computer vision (WACV), Ieee2018, pp. 1451–1460. https://doi.org/10.1109/WACV.2018.00163
20. Xiao, J., Zhao, T., Yao, Y., Yu, Q. & Chen, Y. Context augmentation and feature refinement network for tiny object detection, (2021).
21. Yang, W. et al. Advancing image Understanding in poor visibility environments: A collective benchmark study. *IEEE Trans. Image Process.* **29**, 5737–5752. https://doi.org/10.1109/TIP.2020.2981922 (2020).
22. Ge, Z., Liu, S., Wang, F., Li, Z. & Sun, J. Yolox: Exceeding yolo series in 2021, arXiv preprint arXiv:2107.08430, (2021). https://doi.org/10.48550/arXiv.2107.08430
23. Wang, C. Y., Bochkovskiy, A. & Liao, H. Y. M. YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors, Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition pp. 7464–7475. (2023).
24. Carion, N. et al. End-to-end object detection with transformers, European conference on computer vision, Springer2020, pp. 213–229.
25. Wang, C. Y., Yeh, I. H. & Liao, H. Y. M. YOLOv9: learning what you want to learn using programmable gradient information, arxiv Preprint arxiv:2402.13616, (2024). https://doi.org/10.48550/arXiv.2402.13616
26. Hashmi, K. A., Kallempudi, G., Stricker, D. & Afzal, M. Z. Featenhancer: Enhancing hierarchical features for object detection and beyond under low-light vision, Proceedings of the IEEE/CVF International Conference on Computer Vision pp. 6725–6735. (2023).
27. Yin, X., Yu, Z., Fei, Z., Lv, W. & Gao, X. Pe-yolo: Pyramid enhancement network for dark object detection, International Conference on Artificial Neural Networks, Springer2023, pp. 163–174.
28. Redmon, J. & Farhadi, A. Yolov3: an incremental improvement, arxiv Preprint arxiv:1804.02767, (2018). https://doi.org/10.48550/arXiv.1804.02767
29. Loh, Y. P. & Chan, C. S. Getting to know low-light images with the exclusively dark dataset. *Comput. Vis. Image Underst.* **178**, 30–42. https://doi.org/10.1016/j.cviu.2018.10.010 (2019).
30. Wang, W., Yang, W. & Liu, J. Hla-face: Joint high-low adaptation for low light face detection, Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition2021, pp. 16195–16204.
31. Cui, Z. et al. Multitask aet with orthogonal tangent regularity for dark object detection, Proceedings of the IEEE/CVF International Conference on Computer Vision pp. 2553–2562. (2021).
32. Liu, Z. et al. Swin transformer: Hierarchical vision transformer using shifted windows, Proceedings of the IEEE/CVF international conference on computer vision pp. 10012–10022. (2021).
33. Chen, J. et al. Run, Don't walk: Chasing higher FLOPS for faster neural networks, Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition pp. 12021–12031. (2023).
34. Woo, S. et al. Convnext v2: Co-designing and scaling convnets with masked autoencoders, Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition pp. 16133–16142. (2023).
35. Li, Y. et al. Rethinking vision transformers for mobilenet size and speed, Proceedings of the IEEE/CVF International Conference on Computer Vision pp. 16889–16900. (2023). https://doi.org/10.48550/arXiv.2212.08059
36. Fu, X. et al. A fusion-based enhancing method for weakly illuminated images. *Sig. Process.* **129**, 82–96. https://doi.org/10.1016/j.sigpro.2016.05.031 (2016).
37. Guo, X., Li, Y. & Ling, H. Low-light image enhancement via illumination map Estimation. *IEEE Trans. Image Process.* **26**, 982–993. https://doi.org/10.1109/TIP.2016.2639450 (2016).
38. Zhang, H. et al. Dino: Detr with improved denoising anchor boxes for end-to-end object detection[J]. (2022). arXiv preprint arXiv:2203.03605.
39. Zhu, J. Y., Park, T., Isola, P. & Efros, A. A. Unpaired image-to-image translation using cycle-consistent adversarial networks, Proceedings of the IEEE international conference on computer vision pp. 2223–2232. (2017).
40. Park, T., Efros, A. A., Zhang, R. & Zhu, J. Y. Contrastive learning for unpaired image-to-image translation, Computer Vision–ECCV : 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IX 16, Springer2020, pp. 319–345. (2020).
41. Li, J. et al. DSFD: dual shot face detector, Proceedings of the IEEE/CVF conference on computer vision and pattern recognition pp. 5060–5069. (2019).
42. Ma, L. et al. Toward fast, flexible, and robust low-light image enhancement[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. : 5637–5646. (2022).

Acknowledgements

This work was supported by the Hubei Province Science and Technology Innovation Talents Project (Grant No. 2023DJC070) and supported by Graduate Innovative Fund of Wuhan Institute of Technology (Grant No.

CX2023296; Grant No.CX2023297).

Author contributions

Qingyuan Pan: Methodology, Validation, Writing-original draft, Formal analysis. Wei Huang: Methodology, Validation, Investigation, Writing-original draft. Qiang Liu: Methodology, Validation, Investigation, Writing-original draft.

Declarations

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to W.H.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.