Toward Highly Efficient Semantic-Guided Machine Vision for Low-Light Object Detection

Xin Feng
xfeng@cqut.edu.cn
Junxian Zeng
dave_zjx@stu.cqut.edu.cn
Siping Wang
spwapex@stu.cqut.edu.cn
Zhenwei He*
hzw@cqut.edu.cn

School of Computer Science and Engineering
Chongqing University of Technology
Chongqing, China

Abstract

Detectors trained on well-lit data often experience significant performance degradation when applied to low-light conditions. To address this challenge, low-light enhancement methods are commonly employed to improve detection performance. However, existing human vision-oriented enhancement methods have shown limited effectiveness, which overlooks the semantic information for detection and achieves high computation costs. To overcome these limitations, we introduce a machine vision-oriented highly efficient low-light object detection method with the Efficient semantic-guided Machine Vision-oriented module (EMV). EMV can dynamically adapt to the object detection part based on end-to-end training and emphasize the semantic information for the detection. Besides, by lightening the network for feature decomposition and generating the enhanced image on latent space, EMV is a highly lightweight network for image enhancement, which contains only 27K parameters and achieves high inference speed. Extensive experiments conducted on ExDark and DarkFace datasets demonstrate that our method significantly improves detector performance in low-light environments. Our code is now available at https://github.com/Zeng555/EMV-YOLO.

1 Introduction

With the advancement of deep learning, object detection models have made remarkable progress in the field of computer vision. Current object detection models, including the single-stage YOLO [1, 14, 19, 20, 21, 24] series and the two-stage RCNN [10, 11, 22] series, are typically trained on high-quality image datasets (e.g. COCO [15], Pascal VOC [8]). However, these models often miss detections in challenging low-light scenarios, compromising their reliability. To this end, recently, many works have been proposed to improve

* Corresponding author.

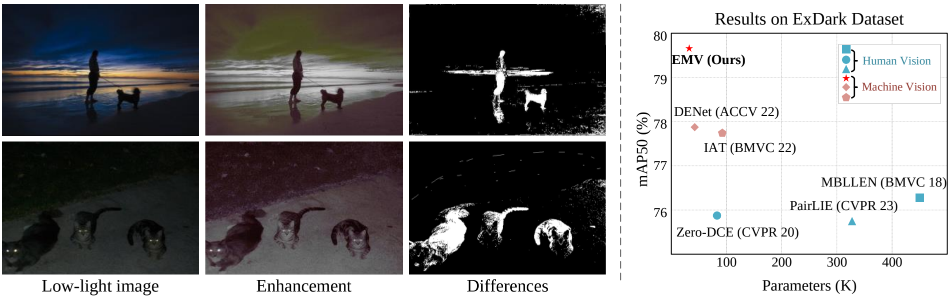

Figure 1: The left part: our method effectively enhances the semantic information of objects while suppressing background. The right part: the comparison results between our method and previous methods on ExDark dataset, our EMV achieves superior performance while containing the fewest parameters.

low-light image visual perception for downstream tasks like object detection, semantic segmentation, and depth estimation.

Generally, low-light object detection can be divided into two frameworks: human vision-oriented framework and machine vision-oriented framework. The human vision-oriented framework is a two-stage approach, where the enhancement network and detector are independent modules. The enhancement network is typically pre-trained on paired low/normal-light image datasets. It first enhances the low-light images to well-lit images before feeding them into the detector for training. On the other hand, the machine vision-oriented framework is the one-stage approach, where the enhancement network and detector are usually end-to-end connected and optimized together. The enhancement network aims to improve the detection results rather than optimizing for the brightness enhancement of the image. However, enhancement networks [12, 17, 25] in human vision-oriented frameworks typically focus on improving image brightness and color balance for a human-friendly effect, often neglecting the visual perception requirements of object detection models. This can degrade semantic information during enhancement, leading to poorer detection of small or occluded objects. Consequently, as noted by researchers in [4, 18], images optimized for human vision may not be ideal for low-light object detection.

Furthermore, some enhancement methods overlook inference speed, causing delays in detection due to the time taken for image enhancement. Most enhancement methods [28, 32] utilize down-sampling and up-sampling techniques, such as Laplacian pyramids, to enhance low-frequency information and restore high-frequency details. These approaches can make the enhancement models overly complex. Even in low-level tasks based on Retinex theory [13], the use of complex decomposition networks for image enhancement [26, 30, 31] further hampers detection speed.

To solve the above problems, in this paper, we propose a machine vision-oriented highly efficient low-light object detection method with the Efficient semantic-guided Machine Vision-oriented module (EMV). EMV can adapt to the object detection part and improve the performance of the model with end-to-end training. Specifically, EMV decomposes low-light images into low-level features in the low-level part and enhances the semantic information in the high-level part.

The low-level task involves decomposing low-light images into two low-level features:

Reflectance and Illumination. Following the visual physics model of the Retinex Theory, Reflectance contains color information, and Illumination contains texture details. Since Retinex decomposition is a low-level task, we devised a lightweight network to execute the decomposition, reducing parameter count while effectively completing the task.

In the high-level part, EMV aims to enhance the semantic information of the image for the detection task. To enhance the semantics of low-light data, both Reflectance and Illumination information are crucial. Reflectance contains color information, while Illumination preserves texture details. By adaptively combining these two features, the enhancement can emphasize object semantics for detection. As shown in the left part of Figure 1, where the differences between the original and enhanced images are highlighted, we observe that the foreground objects are brightened.

Inspired by the Latent Diffusion Model [23], we propose a latent feature enhancement module to enhance the image semantic features in the latent feature space, where enhancement is operated on the low-dimension features rather than the original image, leading to much less computational cost. Furthermore, we do not adopt the up-sampling process when generating the enhanced image, which makes our method much more efficient.

We conducted extensive and fair experiments on ExDark [16] and UG^2+DarkFace [27] and achieved superior results compared to state-of-the-art methods in recent years, our approach has the smallest parameter count, only 27k, making it well-suited for integration into low-light object detection tasks. The experiments in the right part of Figure 1 demonstrate our effectiveness. As we can see, with the minimum number of parameters, the proposed EMV obtains the highest detection accuracy in terms of mAP50 compared with state-of-the-art low-light object detection methods.

Our contributions can be summarized as follows:

- We introduce the Efficient semantic-guided Machine Vision (EMV) module based on the Retinex decomposition network, which adaptively balances the reflectance and illumination information for the detection model.

- To improve the inference speed of EMV, we introduce a lightweight network for feature decomposition, incorporate Latent Feature Enhancement in latent space, and remove the up-sampling phase for image generation. EMV only contains **27K** parameter and exhibits fast processing speed.

- Through extensive and fair experimental comparisons, our method demonstrates superior performance compared to state-of-the-art methods on low-light object detection tasks.

2 Related Work

In low-light object detection, it can be divided into the following three common paradigms.

Human Vision-oriented Approaches. Such methods typically involve using a pretrained low-light image enhancement network to enhance a dataset of low-light images, aiming to restore them as closely as possible to well-lit scenes. Many works [2, 9, 26, 28, 30, 31] is based on Retinex Theory. Wei *et al.* [26] combined the Retinex theory with deep learning, designing a deep network for decomposition and enhancement, and utilized BM3D [6] for image denoising.

Machine Vision-oriented Approaches. In such methods, the enhancement networks are aiming to improve object detection accuracy rather than restoring low-light images to scenes with good lighting as much as possible. Cui *et al*. [5] proposed Illumination Adaptive Transformer (IAT), by employing an inverse mapping function, the sRGB image is transformed into its corresponding raw-RGB space, facilitating dynamic adjustment of image brightness through key parameters in the ISP process, such as gamma values, white balance, and related color matrices.

Domain Adaption Approaches. In low-light object detection, these methods help pre-trained models adjust to new scenarios, improving detection performance in low-light conditions. Du *et al*. [7] introduced DAI-Net, which enhances low-light object detection through day-night domain adaptation, integrating Retinex theory into a reflectance representation learning module and introducing an interchange-redecomposition-coherence procedure to improve image decomposition. However, such methods may encounter challenges in effectively capturing intricate domain variations, thereby resulting in suboptimal performance under specific circumstances.

3 Methodology

3.1 Preliminary

The training strategy in existing human vision-oriented low-light object detection methods involves first through a pre-trained low-light image enhancement model to enhance the low-light images I_{low} to obtain images that closely resembles well-lit conditions (*i.e.*, human visual quality), denoted as I_{high}. Mathematically, it can be expressed as Equation 1,

$$I_{high} = E(I_{low}). \tag{1}$$

where E is the enhancement network, I_{low} are the low-light images in the original dataset, and I_{high} are the images in the enhanced dataset. Subsequently, these enhanced images are used to train the detector, thereby achieving the goal of low-light object detection: $D(I_{high})$.

In the machine-vision-oriented low light object detection paradigm, both the image enhancement and object detection are integrated in the end-to-end manner, which means that the mapping function takes low-light images as input , *i.e.*, $D(I_{low})$, and outputs the overall detection results directly.

3.2 Overview

Our method is designed for machine vision, which can be effectively integrated with detection tasks. The overall framework, as shown in Figure 2, comprises three parts: firstly, Retinex decomposition (green block in Figure 2), also known as the low-level task (see Sec 3.3), which employs a lightweight shallow network to decompose the low-light image S into reflectance component R and illumination component I. Secondly, feature enhancement (light red portion in Figure 2), where the two decomposed components are fed into two separate branches for enhancement, referred to as the high-level task (see Sec 3.4). Finally, the last part involves merging the two enhanced components $[\bar{I}, \bar{R}]$ to form the enhanced low-light image $\bar{S}$. During training, our EMV is end-to-end connected with the YOLOv3 object detector [20], namely EMV-YOLO. The enhanced images $\bar{S}$ are then fed into the detector,

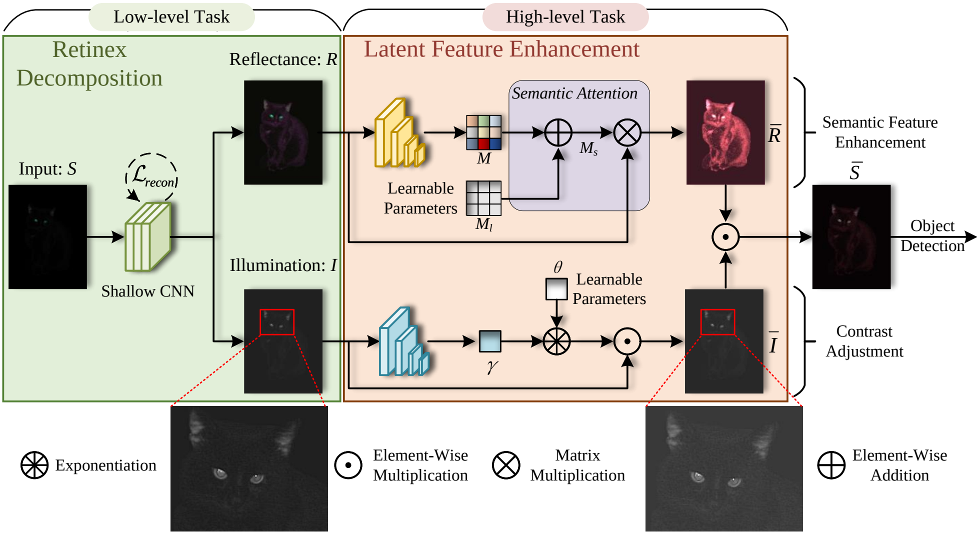

Figure 2: The overview framework of our EMV.

and the gradients calculated from the detection loss are used to simultaneously update the model parameters of both EMV and YOLOv3.

3.3 Retinex Decomposition

Retinex decomposition is considered a relatively low-level vision task, therefore, we utilizing a shallow network to effectively accomplish the decomposition task, $f_{decom}(S) = [I, R]$. R consists of 3 channels, representing the color details of S, while I is a single-channel representation, capturing the texture details of S. The components I and R obtained from Retinex decomposition do not have a standard ground truth. To enable the reconstruction of the original image from I and R (as depicted in the right portion of Figure 4), constraints must be applied to the model, we employ a $\mathcal{L}_2$ loss to constrain the model during the decomposition process.

Mathematically, this is represented as shown in Equation 2,

$$\mathcal{L}_{recon} = min(\sum |S - (I \odot R)|^2). \tag{2}$$

where S represents the input low-light image, while I and R denote the decomposed illumination and reflectance components, respectively.

3.4 Latent Feature Enhancement

To better extract high-level information contained in the two components, we deeply encode the original components to convert them into latent space representations. By reducing the dimensionality of the feature data, we obtain critical abstractions of the original components.

Semantic Feature Enhancement. The reflectance component R enables us to capture the semantic information inherent in the low-light image S, thereby facilitating the enhancement of its semantic features.

Specifically, we apply convolutional layers and down-sampling operations to perform deep feature extraction on the reflectance component R. The resulting feature map is then transformed into a semantic matrix M using a Multilayer Perceptron (MLP), capturing essential semantic information contained within the reflectance component R. This process can be expressed as

$$M = \mathcal{E}(R). \tag{3}$$

$\mathcal{E}$ represents the networks to extract semantic features. Following this, we initialize an identity matrix, which matches the dimensions of the semantic matrix M. To adjust the semantic feature of R adaptively, this identity matrix is transformed into learnable parameters, denoted as M_l. Subsequently, these matrices, along with R, are inputted into our **Semantic Attention** mechanism (purple block of the high-level task in Figure 2) to yield an semantic-enriched reflectance component, $\bar{R}$. This process is depicted as follows:

$$M_s = M \oplus M_l, \quad \bar{R} = M_s \otimes R. \tag{4}$$

where M_s represents a weighted semantic matrix, with its weights being adaptable parameters optimized during training. As illustrated by $\bar{R}$ in Figure 2, this method effectively enhances relevant semantic details in the image, facilitating the distinction between foreground and background elements, thereby achieving the desired enhancement outcome.

Contrast Adjustment. The illumination component I is decomposed as a single-channel grayscale image. At this point, the color information of the original image is lost, leaving only the grayscale information. The removal of color information emphasizes the edge of the image, thereby making the foreground object more prominent and clear.

Our enhancement method focuses not on altering the brightness of the image, but rather on fine-tuning the contrast of specific key features within the original illumination component, denoted as I. Initially, we employ a neural network to extract a contrast factor, γ. The factor indicates the contrast levels across different parts of the image. Mathematically, it can be expressed as $\gamma = \mathcal{E}(I)$. Following this, we introduce a trainable parameter θ and set as the exponent of γ to fine-tune the texture details: $\gamma_t = \gamma^{\theta}$, where γ_t denotes the adjusted contrast factor, this adjustment effectively fine-tunes the texture information. Finally, we combine these features with the original illumination I, thereby adjusting the contrast of the objects in the image: $\bar{I} = \gamma_t \odot I$, where $\bar{I}$ denotes the enhanced illumination.

To enhance both the semantic features of the objects and restore the edge and texture details, we fuse the two enhanced components to reconstruct the low-light enhanced image $\bar{S}$ (as depicted in the left portion of Figure 4). This fusion process is mathematically expressed as $\bar{S} = \bar{I} \odot \bar{R}$.

3.5 Network Training

During training, our EMV shares the detection loss with the detector. Additionally, our low-level decomposition task is constrained by a simple extra loss, ensuring that the decomposed R and I can effectively reconstruct S.

Therefore, the total loss for our low-light object detection consists of two parts: Retinex reconstruction loss, denoted as $\mathcal{L}_{recon}$; YOLOv3 object detecion loss, denoted as $\mathcal{L}_{detect}$. The detecion loss is implemented within MMDetection [3]. Mathematically, total loss used for training can be represented as

$$\mathcal{L} = \mathcal{L}_{detect} + \lambda_{recon} \cdot \mathcal{L}_{recon}. \tag{5}$$

λ_{recon} denote the hyper-parameter assigned to our $\mathcal{L}_{recon}$.

4 Experiments

In this section, we conducted two main experiments: low-light object detection (Sec. 4.1) and low-light face detection (Sec. 4.2). Furthermore, we conducted several ablation experiments (Sec. 4.3) to evaluate the effectiveness of the proposed modules. Finally, we also performed efficiency analysis experiments (Sec. 4.4). Our experiments were conducted using the object detection framework MMDetection [3].

4.1 Low-light Object Detection

Hyper-parameters Settings. This training process fine-tuned a pretrained detection model on the COCO dataset using the SGD optimizer for 25 epochs. The model was trained with a batch size of 12, a momentum of 0.9, and a weight decay of 4e-5. A learning rate warm-up schedule was applied for the first 2,000 iterations, gradually increasing the learning rate to 1e-3, and was then decayed by a factor of 0.1 at the 6th, 11th, and 16th epochs. Loss weight λ_{recon} in Equation 5 is set to 500.

Dataset. The proposed framework was evaluated on the ExDark dataset, which is a commonly used real-world dataset for low-light object detection. ExDark dataset consists of 12 object categories and 7,363 low-light images. To maintain consistency with previous research in the field [4, 5, 29], we followed a similar data split strategy, allocating 80% of each category for training and the remaining 20% for evaluation purposes.

The experimental results are presented in Table 1, our EMV-YOLO method outperforms other state-of-the-art methods. Approaches such as [9, 12, 17, 20, 30] are human vision-oriented approaches, while [7] follows a domain adaptation approach. The remaining approaches, [4, 5, 18, 29], are machine vision-oriented. From Table 1, we can observe that machine vision-based methods generally outperform human vision-based methods. Our EMV-YOLO achieved the highest mAP among various compared methods, surpassing the state-of-the-art DAI-Net [7] by 1.4 points. The visualized results can be seen from Figure 3, in the low-light scene, our method detected all the objects in the Ground Truth, while such as methods in [5, 9, 12, 17, 18, 29, 30], missed some objects. This demonstrates the superiority of our method.

Table 1: Detection evaluation on ExDark dataset, the red font indicates the best performance, while the blue font indicates the second-best performance.

Method	Bicycle	Boat	Bottle	Bus	Car	Cat	Chair	Cup	Dog	Motorbike	People	Table	mAP50(%)
YOLOv3	79.8	75.3	78.1	92.3	83.0	68.0	69.0	79.0	78.0	77.3	81.5	55.5	76.4
MBLLEN [17] + YOLOv3	81.9	76.6	78.2	91.1	**84.5**	69.3	69.0	78.3	77.8	73.3	81.5	54.0	76.3
KinD [30] + YOLOv3	80.9	75.0	75.8	93.3	82.4	69.4	69.2	79.0	76.9	76.3	79.6	55.4	76.1
Zero-DCE [12] + YOLOv3	81.2	75.0	75.7	93.4	83.2	67.7	70.2	76.4	74.1	77.7	81.3	55.5	75.9
PairLIE [9] + YOLOv3	80.8	78.3	76.8	90.5	**84.5**	66.8	69.1	75.6	78.9	73.7	80.3	54.5	75.8
DAI-Net [7]	**83.8**	75.8	75.1	**94.2**	84.1	**74.9**	**73.1**	79.2	**82.2**	76.4	80.7	**59.8**	**78.3**
MAET [4]	83.1	78.5	75.6	92.9	83.1	73.4	71.3	79.0	79.8	77.2	81.1	57.0	77.7
DENet [18]	80.9	**79.2**	**80.1**	90.7	**84.5**	70.7	72.0	79.3	80.1	76.7	**82.4**	58.0	77.9
PE-YOLO [29]	**84.7**	**79.2**	79.3	92.5	83.9	71.5	71.7	**79.7**	79.7	77.3	81.8	55.3	78.0
IAT-YOLO [5]	79.8	76.9	78.6	92.5	83.8	73.6	72.4	78.6	79.0	**79.0**	81.1	57.7	77.8
EMV-YOLO (ours)	82.8	**79.7**	**79.8**	**94.1**	**84.7**	**74.3**	**74.1**	**83.1**	**82.7**	**78.1**	**83.6**	**59.3**	**79.7**

4.2 Low-light Face Detection

Hyper-parameters Settings. The main hyper-parameter settings remain consistent with previous section (Sec. 4.1); the only difference is the learning rate decay strategy, which now occurs at the 8th and 16th epochs.

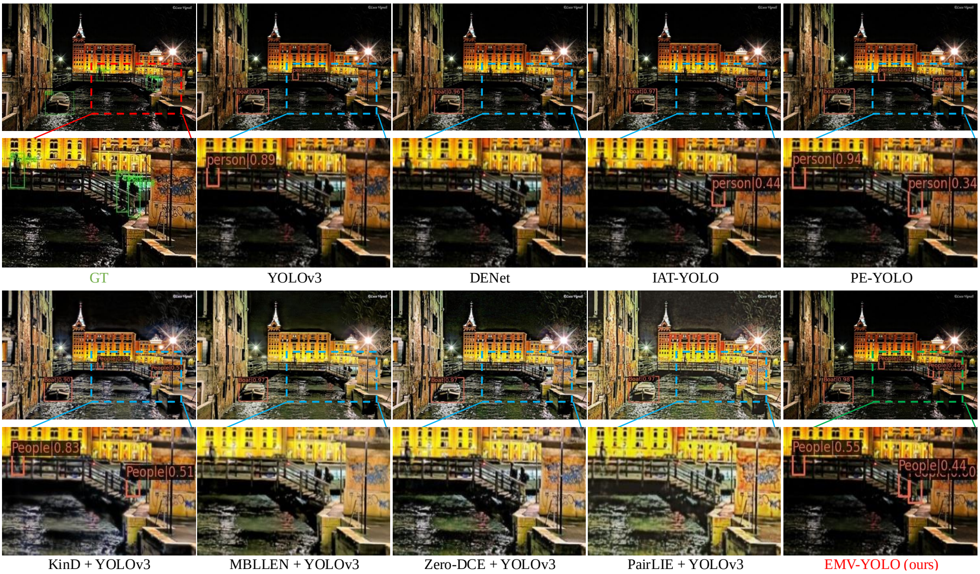

Figure 3: The visualized results on the ExDark dataset demonstrate that our EMV-YOLO model surpasses other methods in detecting objects in low-light scenes.

9Method	Type	mAP50(%)
YOLOv3	baseline	48.3
MBLLEN [17] + YOLOv3	Human Vison	51.6
KinD [30] + YOLOv3		51.6
Zero-DCE [12] + YOLOv3		54.2
PairLIE [9] + YOLOv3		55.4
DAI-Net [7]	Domain Adaptation	57.0
MAET [4]	Machine Vision	55.8
DENet [18]		51.2
PE-YOLO		51.1
IAT-YOLO [5]		53.1
EMV-YOLO (ours)		**57.6**

Table 2: Detection evaluation on DarkFace dataset.

Method	$\mathcal{L}_{recon}$	λ_{recon}	ExDark	UG^2+DarkFace
YOLOv3	×	×	76.4	48.3
EMV-YOLO (ours)	✓	0	78.2	53.3
	✓	300	**79.4**	53.6
	✓	400	**79.4**	52.7
	✓	500	**79.7**	**57.6**
	✓	600	79.1	**56.2**

Table 3: Ablation study for our method with different loss weight λ_{recon}.

Dataset. We conduct evaluation experiments on UG^2+DarkFace dataset. DarkFace is a low-light face detection dataset in real-world scenarios, comprising 6,000 images. We divided the images into a training set of 5,400 images and a testing set of 600 images, following the division approach used [4].

We compared our method with the same methods as in previous section (Sec. 4.1). As shown in the Table 2, our approach continues to achieve excellent results, outperforming the state-of-the-art method DAI-Net [7] by 0.6 points.

4.3 Ablation Study

Optimal Loss Weights. This study aims to determine the optimal weight for the $\mathcal{L}_{recon}$ loss used to rectify the Retinex decomposition. As indicated in Table 3, the best results were achieved when λ_{recon}=500 on both datasets, we hence adopted λ_{recon}=500 as the optimal weight for the $\mathcal{L}_{recon}$ loss.

Retinex Decomposition Encoder. This study aims to assess the difference between our decomposition module (the module in the green block in Figure 2) and the decomposition net used in RetinexNet [26]. During experiments, we individually tested these two different decomposition modules in the end-to-end machine vision-oriented object detection framework. As demonstrated in Table 4, when the framework utilized our proposed encoder, it achieved superior performance with model parameter of only **27k**. In contrast, the decomposition net of [26] resulted in a parameter count exceeding 8 times more, and with less improvement in detection accuracy. This result confirms the efficacy of our decomposition encoder.

Table 4: Ablation study on different Retinex Decompostion Net.

Method	RetinexNet [26]	Ours	Parameters (K)	ExDark	UG^2+DarkFace
YOLOv3	×	×	×	76.4	48.3
EMV-YOLO (ours)	✓	×	225.3	76.6	52.8
	×	✓	**27**	**79.7**	**57.6**

Feature Enhancement Modules. This study aims to evaluate the effectiveness of our feature enhancement modules. Our Retinex decomposition encoder is used by default, where the decomposition is performed prior to the enhancement. The module tailored for enhancing the illumination component is termed as Illumination Enhancement Module (IEM), while the module dedicated to enhancing the reflectance component is referred to as Reflectance Enhancement Module (REM). In our experiments, we evaluated three configurations: (i) neither module was used to enhance the original I and R components; (ii) only one of the modules was employed to enhance either the I or R component; (iii) both modules were utilized concurrently to enhance the I and R components. As shown in Table 5, compared to the baseline [20], use each of the single module could improve the mAP50 accuracy on both ExDark and UG^2+DarkFace datasets, respectively, while the combined usage of both modules yielded the best performance. The same outcome can be seen from the visualization results shown in the left part of Figure 4 as well.

Table 5: Ablation study on each enhancement module of the our model.

Method	IEM	REM	Parameters (K)	ExDark	UG^2+DarkFace
YOLOv3	×	×	×	76.4	48.3
EMV-YOLO (ours)	×	×	9.6	78.7	52.4
	✓	×	16	79.1	54.1
	×	✓	20	78.9	53.7
	✓	✓	27	**79.7**	**57.6**

Table 6: Efficiency analysis.

Method	Type	Parameters	Runtime (ms)	FLOPs (G)
MBLLEN [17]	Human Vision	450K	38.1	81.98
KinD [30]		8M	14.8	54.17
Zero-DCE [12]		79K	4.7	23.44
PairLIE [9]		342K	13.6	100.91
DENet [18]	Machine Vision	**45K**	**3.3**	**1.42**
PE-YOLO [29]		91K	32.2	8.11
IAT-YOLO [6]		91K	13.8	6.48
EMV-YOLO (ours)		**27K**	**4.1**	**5.17**

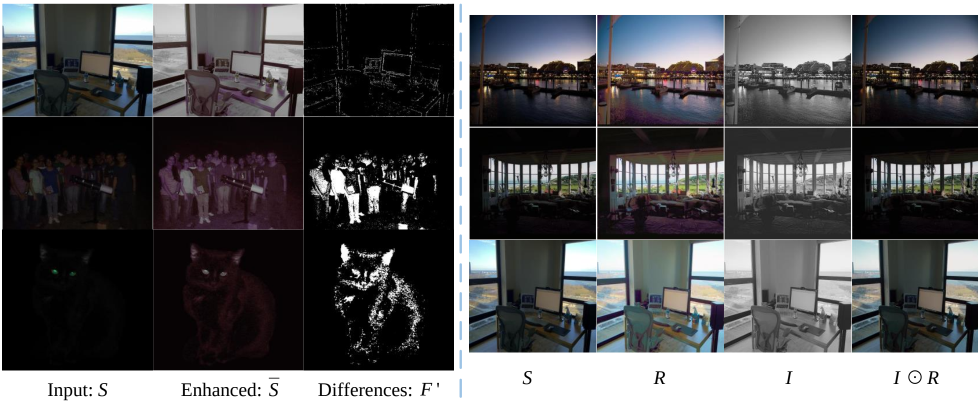

Figure 4: The left portion illustrates the enhancement results through the latent feature enhancement. The right portion demonstrate the decomposed components I and R that are trained from $\mathcal{L}_{recon}$ constraint can be well reconstructed as the original low-light image S.

4.4 Efficiency Analysis

To investigate the parameters, inference time (measured in ms), and computational complexity (measured in GFLOPs) of different methods, we evaluate the efficiency of comparison models with a 3-channel, 544×544-pixel input image on a single RTX 3090 GPU. As can be seen from Table 6, our proposed EMV-YOLO exhibits the lowest parameter amount and the second lowest runtime and computational load among all compared methods, demonstrating our superior performance in efficiency.

5 Conclusion

In this paper, we propose a Efficient semantic-guided Machine Vision-oriented module for low-light object detection, namely EMV. Toward improving the efficiency and performance of machine vision-oriented low-light object detection, this module decomposes the low-light images by using a lightweight decomposition encoder based on Reintex theory, and further enhances the objects' semantic information and texture details in a latent feature enhancement process, this enhances the semantic information of objects while suppressing background. EMV is embedded and optimized in the end-to-end object detection framework. Extensive experiments on both the low-light object and face detection tasks demonstrate that the proposed EMV-YOLO outperforms state-of-the-art methods in terms of both detection accuracy and model complexity.

Acknowledgement

This work is supported by the Key Project of Chongqing Technology Innovation and Application Development (Grant No. cstc2021jscx-dxwtBX0018), the Natural Science Foundation of Chongqing (Grant No. CSTB2022NSCQ-MSX0493), National Natural Science Foundation of China (Grant No. 62306053), the Graduate Innovation Project of Chongqing University of Technology (Grant No. gzlcx20233250).

References

[1] Alexey Bochkovskiy, Chien-Yao Wang, and Hong-Yuan Mark Liao. Yolov4: Optimal speed and accuracy of object detection. *arXiv preprint arXiv:2004.10934*, 2020.

[2] Yuanhao Cai, Hao Bian, Jing Lin, Haoqian Wang, Radu Timofte, and Yulun Zhang. Retinexformer: One-stage retinex-based transformer for low-light image enhancement. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 12504–12513, 2023.

[3] Kai Chen, Jiaqi Wang, Jiangmiao Pang, Yuhang Cao, Yu Xiong, Xiaoxiao Li, Shuyang Sun, Wansen Feng, Ziwei Liu, Jiarui Xu, Zheng Zhang, Dazhi Cheng, Chenchen Zhu, Tianheng Cheng, Qijie Zhao, Buyu Li, Xin Lu, Rui Zhu, Yue Wu, Jifeng Dai, Jingdong Wang, Jianping Shi, Wanli Ouyang, Chen Change Loy, and Dahua Lin. MMDetection: Open mmlab detection toolbox and benchmark. *arXiv preprint arXiv:1906.07155*, 2019.

[4] 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 (ICCV)*, pages 2553–2562, 2021.

[5] Ziteng Cui, Kunchang Li, Lin Gu, Shenghan Su, Peng Gao, ZhengKai Jiang, Yu Qiao, and Tatsuya Harada. You only need 90k parameters to adapt light: a light weight transformer for image enhancement and exposure correction. In *33rd British Machine Vision Conference 2022, BMVC 2022, London, UK, November 21-24, 2022*. BMVA Press, 2022.

[6] Kostadin Dabov, Alessandro Foi, Vladimir Katkovnik, and Karen Egiazarian. Image denoising with block-matching and 3d filtering. In *Image processing: algorithms and systems, neural networks, and machine learning*, volume 6064, pages 354–365. SPIE, 2006.

[7] 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*, pages 12666–12676, 2024.

[8] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International Journal of Computer Vision*, 88:303–338, 2010.

[9] 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*, pages 22252–22261, 2023.

[10] Ross Girshick. Fast r-cnn. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 1440–1448, 2015.

[11] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the*

IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 580–587, 2014.

[12] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1780–1789, 2020.

[13] Edwin H Land. The retinex theory of color vision. *Scientific american*, 237(6):108–129, 1977.

[14] Chuyi Li, Lulu Li, Hongliang Jiang, Kaiheng Weng, Yifei Geng, Liang Li, Zaidan Ke, Qingyuan Li, Meng Cheng, Weiqiang Nie, et al. Yolov6: A single-stage object detection framework for industrial applications. *arXiv preprint arXiv:2209.02976*, 2022.

[15] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014.

[16] 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.

[17] Feifan Lv, Feng Lu, Jianhua Wu, and Chongsoon Lim. MBLLEN: low-light image/video enhancement using cnns. In *British Machine Vision Conference 2018, BMVC 2018, Newcastle, UK, September 3-6, 2018*, page 220. BMVA Press, 2018.

[18] Qingpao Qin, Kan Chang, Mengyuan Huang, and Guiqing Li. Denet: Detection-driven enhancement network for object detection under adverse weather conditions. In *Proceedings of the Asian Conference on Computer Vision (ACCV)*, pages 2813–2829, 2022.

[19] Joseph Redmon and Ali Farhadi. Yolo9000: better, faster, stronger. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7263–7271, 2017.

[20] Joseph Redmon and Ali Farhadi. Yolov3: An incremental improvement. *arXiv preprint arXiv:1804.02767*, 2018.

[21] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 779–788, 2016.

[22] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28, 2015.

[23] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10684–10695, 2022.

[24] Chien-Yao Wang, Alexey Bochkovskiy, and Hong-Yuan Mark Liao. Yolov7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7464–7475, 2023.

[25] Yufei Wang, Renjie Wan, Wenhan Yang, Haoliang Li, Lap Pui Chau, and Alex Kot. Low-light image enhancement with normalizing flow. In *36th AAAI Conference on Artificial Intelligence, AAAI 2022*, pages 2604–2612. Association for the Advancement of Artificial Intelligence, 2022.

[26] Chen Wei, Wenjing Wang, Wenhan Yang, and Jiaying Liu. Deep retinex decomposition for low-light enhancement. In *British Machine Vision Conference 2018, BMVC 2018, Newcastle, UK, September 3-6, 2018*, page 155. BMVA Press, 2018.

[27] Wenhan Yang, Ye Yuan, Wenqi Ren, Jiaying Liu, Walter J Scheirer, Zhangyang Wang, Taiheng Zhang, Qiaoyong Zhong, Di Xie, Shiliang Pu, et al. Advancing image understanding in poor visibility environments: A collective benchmark study. *IEEE Transactions on Image Processing*, 29:5737–5752, 2020.

[28] Xunpeng Yi, Han Xu, Hao Zhang, Linfeng Tang, and Jiayi Ma. Diff-retinex: Rethinking low-light image enhancement with a generative diffusion model. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 12302–12311, 2023.

[29] Xiangchen Yin, Zhenda Yu, Zetao Fei, Wenjun Lv, and Xin Gao. Pe-yolo: Pyramid enhancement network for dark object detection. In *International Conference on Artificial Neural Networks*, pages 163–174. Springer, 2023.

[30] Yonghua Zhang, Jiawan Zhang, and Xiaojie Guo. Kindling the darkness: A practical low-light image enhancer. In *Proceedings of the 27th ACM international conference on multimedia*, pages 1632–1640, 2019.

[31] Yonghua Zhang, Xiaojie Guo, Jiayi Ma, Wei Liu, and Jiawan Zhang. Beyond brightening low-light images. *International Journal of Computer Vision*, 129:1013–1037, 2021.

[32] Dewei Zhou, Zongxin Yang, and Yi Yang. Pyramid diffusion models for low-light image enhancement. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence, IJCAI 2023, 19th-25th August 2023, Macao, SAR, China*, pages 1795–1803. ijcai.org, 2023.