Trash to Treasure: Low-Light Object Detection via Decomposition-and-Aggregation

Xiaohan Cui[1], **Long Ma**[1], **Tengyu Ma**[1], **Jinyuan Liu**[2], **Xin Fan**[1], **Risheng Liu**[1*]
[1]School of Software Technology, Dalian University of Technology
[2]School of Mechanical Engineering, Dalian University of Technology
malone94319@gmail.com, atlantis918@hotmail.com, {cuixiaohan1230, matengyu}@mail.dlut.edu.cn, {xin.fan, rsliu}@dlut.edu.cn

Abstract

Object detection in low-light scenarios has attracted much attention in the past few years. A mainstream and representative scheme introduces enhancers as the pre-processing for regular detectors. However, because of the disparity in task objectives between the enhancer and detector, this paradigm cannot shine at its best ability. In this work, we try to *arouse the potential of enhancer + detector*. Different from existing works, we extend the illumination-based enhancers (our newly designed or existing) as a scene decomposition module, whose removed illumination is exploited as the auxiliary in the detector for extracting detection-friendly features. A semantic aggregation module is further established for integrating multi-scale scene-related semantic information in the context space. Actually, our built scheme successfully transforms the "*trash*" (i.e., the ignored illumination in the detector) into the "*treasure*" for the detector. Plenty of experiments are conducted to reveal our superiority against other state-of-the-art methods. The code will be public if it is accepted.

Introduction

Object detection is a representative, familiar vision task both in industrial and academic communities. Along with the development of deep learning techniques (Liu et al. 2023a,b), object detection in the regular environment has achieved prominent achievements. However, in specific settings, such as salient object detection (Piao et al. 2019, 2020; Zhang et al. 2020) and low-light object detection, there remains immense potential for research and optimization. Especially, images taken in low light conditions (e.g., nighttime) usually contain complex degradation to heavily limit the performance of regular detectors (He et al. 2016; Jin et al. 2021; Tang et al. 2020; Liu et al. 2021d; Tang et al. 2022; Liu et al. 2021a, 2020). This degradation can be due to factors such as noise, low contrast, and insufficient illumination, which can reduce the visibility of objects and make them hard to detect. In the following, we will review existing works and introduce our contributions.

Related Works

In the past few decades, various schemes have emerged for handling low-light object detection, which can be roughly divided into two categories. The mainstream one is to cascade enhancer and detector, the other is to specifically-design the low-light detector.

For the enhancer-introduced schemes, the well-known UG2+ Challenge competition is a landmark event to drive research for low-light face detection. The cascade of enhancement and detection is the most common solution in the two consecutive championship schemes, e.g., the CAS-Newcastle team adopted this cascaded scheme (Yuan et al. 2019), which illustrates the significance of researching the cascaded pattern. In addition, the latest enhancement schemes (e.g., (Guo et al. 2020; Liu et al. 2022a; Ma et al. 2022b; Xue et al. 2022; Ma et al. 2023, 2021, 2022a)), and the work (Lv, Li, and Lu 2021) are no longer satisfied with the improvement of visual quality and try to apply the well-designed enhancer to the detection task to verify the effectiveness of the algorithm. For example, the improvement of detection performance on the Dark Face (Yang et al. 2020b; Liu et al. 2021b) dataset is a crucial experiment to demonstrate enhanced performance. However, these works often bring limited performance improvement, and it is even better to directly detect low light images. An intuitive explanation is that these enhancements are designed for human-friendly visual quality (Lore, Akintayo, and Sarkar 2017; Wei et al. 2018; Chen et al. 2018; Li et al. 2018), and are difficult to apply to machine vision tasks focused on high-level semantic information, e.g., object detection. Certain methods produce black edges, retain dark noisy areas, or enhance contrast to improve the overall visual quality, but these enhancements can have a negative impact on the performance of object detection.

The other type of scheme makes an effort to start from the perspective of designing detectors in an end-to-end manner. Through a bidirectional low-level adaptation and multi-task high-level adaptation scheme, HLA (Wang, Yang, and Liu 2021; Wang et al. 2022) proposed a joint high-low adaptation framework. This method converted a face detector trained under normal light into a face detector under low light. While domain adaptation methods can mitigate the issue of having a limited number of labeled datasets, their performance improvements were often modest. The method proposed in REG (Liang et al. 2021) is to seamlessly couple a cycle exposure generation module and a multiple exposure detection module, which improved the detection effects by

---

[*] Corresponding author.

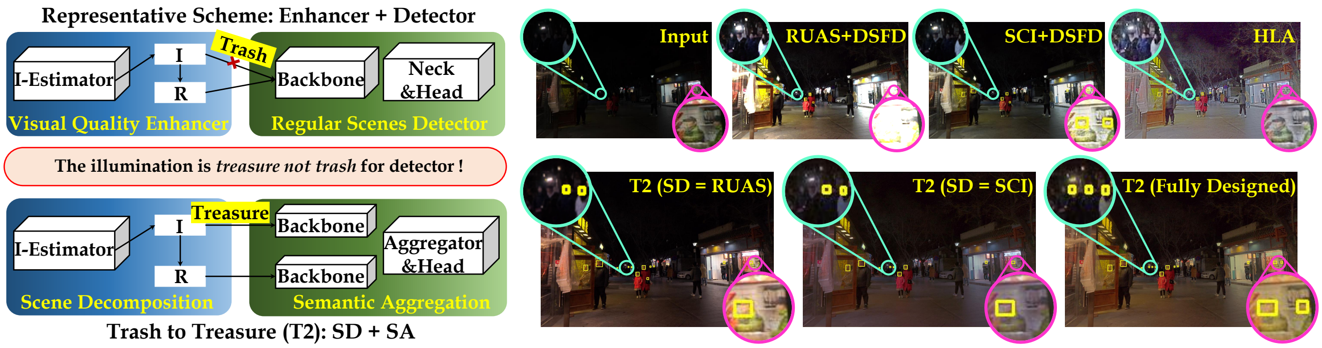


Figure 1: Comparing algorithmic pipelines (*left*) and visual results (*right*). Different from the representative scheme (i.e., enhancer + detector) of low-light detection, we treat the illumination which is viewed as the trash as the treasure, to establish our decomposition-and-aggregation process. It needs to be emphasized that our defined scene decomposition can be instantiated as any illumination-based enhancer. The results in the top row of the right part are two methods based on the representative scheme (enhancer is defined as RUAS and SCI, and the detector is set as DSFD), and a recent method (HLA), respectively. The results in the bottom row are generated by our T2 with different settings of scene decomposition. We can observe that our T2 significantly upgrades the detection ability against the representative scheme. Additionally, our method brightens images (without any loss constraint for enhancement) to realize the enhancement exactly wanted for the detector.

effectively suppressing uneven illumination and noise problems. Moreover, the work (Cui et al. 2021) did not apply the enhancement directly to the low-light image, but used the traditional camera signal processing method to transform the normal-light image into a low-light image, and utilized a predictive transform decoder to predict the parameters involved in the illumination transformation to complete the self-supervised training. (Liu et al. 2022b) proposed a method called IA-YOLO that improves object detection performance in adverse weather conditions. Moreover, clustering helps in extracting better features (Liu et al. 2012; Wu, Lin, and Zha 2019). However, these specifically-designed methods lack the full exploration of scene information, resulting in limited performance improvement.

Our Contributions

To overcome the above drawbacks and arouse the potential of enhancer + detector, this work establishes a new detector with decomposition-and-aggregation by fully exploiting the illumination that is viewed as the trash in the previous schemes. As shown in the left part of Figure 1, we compare the algorithmic pipeline with the representative cascaded scheme. Our proposed **T**rash to **T**reasure (**T2**) is acquired around the fact that "the illumination is a treasure not trash for the detector", to make full use of the complete scene information. Benefiting from the flexibility of our designed scene decomposition, existing illumination-based enhancers can be plugged into T2. The right part in Figure 1 demonstrates the visual results among different state-of-the-art methods. It can be easily observed that T2 not only acquires the best detection accuracy but also significantly ameliorate the detection accuracy for existing methods. Our main contributions can be summarized as

- By deeply analyzing the latent relationship between enhancer and detector, we conclude two key challenges for the cascaded pattern. The one challenge is how to realize detection-oriented enhancement, instead of pure human eye-friendly visual quality. The other is how to reduce information discrepancy between enhanced output and regular data as much as possible.
- By rethinking illumination-based enhancers, we construct a scene decomposition module to acquire two scene-related components for characterizing the low-light scenarios. It supports learning detection-oriented enhancement without introducing additional training constraints related to visual quality.
- To effectively utilize the decomposed components, we design a semantic aggregation module that is composed of a weight-sharing extractor and a multi-scale aggregator. It fully exploits the scene-related content in the feature space to reduce the information discrepancy between regular and low-light data.
- Extensive experimental evaluations are performed to verify our superiority in detection accuracy against other state-of-the-art methods. A series of algorithmic analyses indicate that our T2 successfully realizes the intended target, i.e., detection-oriented enhancement.

Proposed Method

In this section, we present our motivation by analyzing the latent relationship between enhancer and detector. We then construct a scene decomposition module to endow the perception-related ability to the Retinex-based enhancer. Finally, we introduce a semantic aggregation module to strengthen the feature representation. The overall architecture of our proposed method can be found in Figure 2.

Two Key Challenges of Enhancer + Detector

Unlike regular object detection in normal-light environments, the main challenge for low-light object detection is

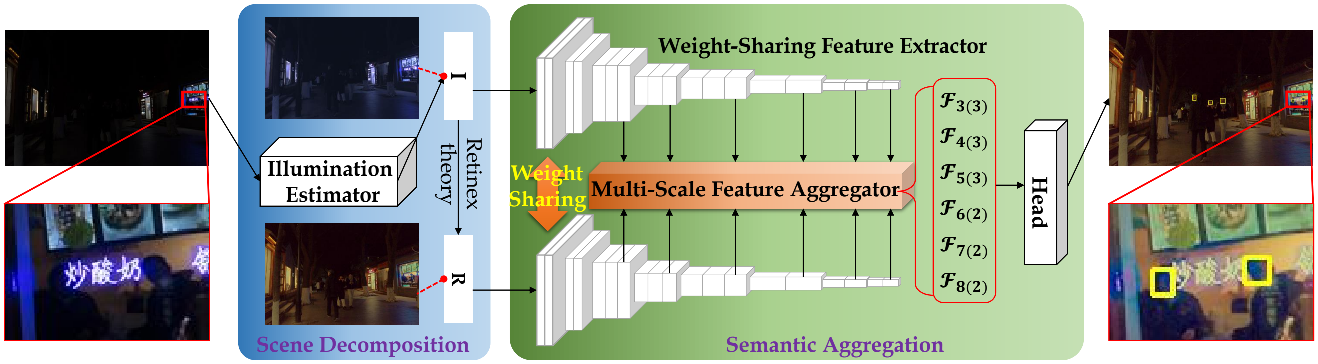


Figure 2: The overall architecture of our proposed method. Our method mainly contains two parts, Scene Decomposition (SD) for generating illumination and reflectance from low-light observation based on Retinex theory, and Scene Aggregation (SA) for integrating multi-scale scene-related information to strengthen feature representation.

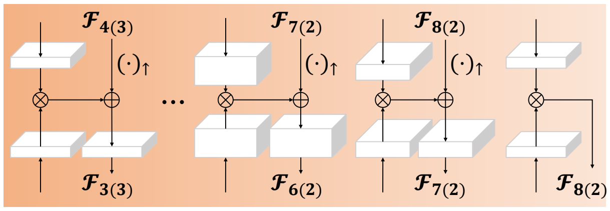


Figure 3: The computational flow of our proposed multi-scale feature aggregator.

that degraded observations with poor visibility heavily influence feature extraction, causing the accuracy to drop sharply. It is a direct and commonly-used manner to generate new visible data by improving visual quality. It is widely adopted in multiple champion solutions born in the UG2+ Challenge competition (landmark for low-light face detection). Among them, experimental explorations suggest that adopting a classical (e.g., MSRCR (Jobson, Rahman, and Woodell 1997)) one with poor visual quality is more effective than the latest advanced enhancer. Therefore, the preconceived conclusion "higher visual quality is more beneficial for detection" cannot be satisfied.

Investigating its reason, on the one hand, these two tasks have different objectives, i.e., pixel-level visual-friendly to human eyes (enhancer) and semantic-level perception-friendly to machines (detector). In other words, the directly cascaded pattern aims at improving detection accuracy, and the enhancer is designed for catering to visual quality, rather than detection-desired high-quality data. On the other hand, the enhancer indeed ameliorates visual quality but it inevitably destroys the inherent distribution that keeps the same situation as natural images, causing the enhanced data to keep a distinct information discrepancy with regularly captured data (maybe low-light, maybe normal-light), which heavily restricts the feature extraction. Combing the above analyses, we can conclude two key challenges for the paradigm of enhancer + detector, described as

1. *What kind of data acquired from the enhancer is required for the detector in low-light scenes?*
2. *How to narrow down the information discrepancy between enhanced results and regular data?*

To settle these two challenges, in the following, we build a scene decomposition module to enable the Retinex-based enhancer beyond visual quality to generate detection-desired data. Then we construct a semantic aggregation module to fully exploit decomposed scene-related features to reduce the information discrepancy.

Scene Decomposition Module

In this part, by rethinking enhancement for detection, we explain the necessity of utilizing the removed illumination. Substantially, we establish a scene decomposition module to acquire the decomposed components.

Rethinking Illumination-based Enhancers Existing low-light image enhancement techniques (Ma et al. 2022b; Liu et al. 2021c) are mostly developed according to the Retinex theory (Land and McCann 1971) (formulated as $\mathbf{L} = \mathbf{R} \otimes \mathbf{I}$, where $\otimes$ denotes the element-wise multiplication), this principle describes that low-light observation $\mathbf{L}$ can be decomposed as the normal-light image $\mathbf{R}$ (also called reflectance) and the removed illumination $\mathbf{I}$. We can find a basic fact from this model, i.e., the enhanced result needs to remove the illumination from the low-light observation. This means the enhanced result cannot keep the same information capacity as the original observation and the missing content is exactly the removed illumination based on the information conservation. However, most detectors are established in regular data without information reduction.

From this perspective, the illumination is abandoned for the enhancer (only needs to focus on the normal-light image), but it should be utilized for the detector to improve the information utilization to the maximum extent.

Decomposing Scene for Detection As described above, we know the removed illumination is essential for detection. That is to say, the two decomposed components from low-light observations are equally important for detection, rather

Method	**Description**	**mAP(%)**
RUAS	Fine-tune detector	65.98
RUAS^{+}	Jointly training	65.70
RUAS^{++}	T2 (SD = RUAS)	$67.36_{\uparrow\mathbf{1.66}}$
SCI	Fine-tune detector	65.68
SCI^{+}	Jointly training	65.42
SCI^{++}	T2 (SD = SCI)	$68.33_{\uparrow\mathbf{2.91}}$

Table 1: Quantitative results among different versions for illumination-based enhancers including RUAS and SCI.

than only focusing on the visual quality of the single component in the enhancer. Generally, the low-light observations reflect the scene information including objects and background. After performing Retinex theory, these two components still contain the scene information. Therefore, we call the module for generating two decomposed components as Scene Decomposition Module (SDM).

Here we provide a simple setting for SDM, which consists of three residual blocks and each residual block includes two Conv-BN-ReLU layers. The 1×1 convolutional layer is used to adjust the number of channels of the feature map at the time of input and output. Notably, we would like to emphasize that SDM can be initialized by existing Retinex-based enhancer (Please refer to Sec.Discussion for experimental supports).

Semantic Aggregation Module

The previous section has generated two decomposed components by scene decomposition, the next issue is how to exploit them for the detector. Here we build a semantic aggregation module that consists of a weigh-sharing feature extractor and a multi-scale feature aggregator.

Weight-Sharing Feature Extractor Our defined scene decomposition module aims at generating two decomposed components to perform scene information. Although they are acquired based on the knowledge from low-light image enhancement, their output status should be related to the detection accuracy. Here we adopt an VGG-16 used in DFSD (Li et al. 2019) and S3FD (Zhang et al. 2017) as its basic architecture. We use the weight-sharing backbone network to extract the features of the two decomposed components.

Multi-Scale Feature Aggregator We know that the feature pyramid network (Lin et al. 2017) is a commonly-used structure for the detector, which can improve the detection accuracy, especially the extremely small objects. Through our above-built weight-sharing feature extractor, we can obtain two groups of multi-scale scene-related semantic features in the context space. To effectively integrate them, we define a multi-scale feature aggregator by introducing the Retinex knowledge into the feature pyramid network. This process can be formulated as

$$\begin{cases}\mathcal{F}_{8(2)} = \mathcal{F}^{\mathbf{I}}_{8(2)} \otimes \mathcal{F}^{\mathbf{R}}_{8(2)},\\ \mathcal{F}_{a(b)} = \mathcal{F}^{\mathbf{I}}_{a(b_a)} \otimes \mathcal{F}^{\mathbf{R}}_{a(b_a)} + (\mathcal{F}_{a+1(b_{a+1})})_{\uparrow},\end{cases} \tag{1}$$

where $3 \leq a \leq 7, b \in \{2, 3\}$. $\mathcal{F}^{\mathbf{I}}$ and $\mathcal{F}^{\mathbf{R}}$ represent the generated features from the feature extractor according to illumination and reflectance, respectively. $\mathcal{F}_{a(b)}$ denotes the feature generated in b-th convolutional layer in the a-th convolution block. We are utilizing six layers of features: $\mathcal{F}_{3(3)}$, $\mathcal{F}_{4(3)}$, $\mathcal{F}_{5(3)}$, $\mathcal{F}_{6(2)}$, $\mathcal{F}_{7(2)}$, and $\mathcal{F}_{8(2)}$. $(\cdot)_{\uparrow}$ represents the upsamling operation. The aggregated operation is defined as the element-wise multiplication $\otimes$. It is because multiplication exactly corresponds to the division used in the scene decomposition. In other words, this way reconstructs the original scene information that existed in the original low-light observation. The detailed computational process can be found in Figure 3.

Training Loss

In the training phase, we adopt a multi-task loss function (Liu et al. 2016; Girshick 2015) for the overall network. The objective loss function consists of a weighted sum of location loss $\mathcal{L}_{loc}$ and a confidence loss $\mathcal{L}_{conf}$, formulated as

$$\mathcal{L}(x, c, l, g) = \frac{1}{N}(\mathcal{L}_{conf}(x, c) + \alpha \mathcal{L}_{loc}(x, l, g)), \tag{2}$$

where $\mathcal{L}_{conf}$ is focal loss and $\mathcal{L}_{loc}$ is smooth L1 loss, α represents the trade-off parameter between the two losses. N is the number of the default boxes, $x_{ij} = \{1, 0\}$ represents an indicator that the i-th default box matches the j-th ground truth box, c represents the predicted class confidence scores and l, g denote the predicted box and the ground truth box, respectively. In this paper, we set α to 1.0. We do not apply any loss constraints specifically to the scene decomposition module.

Discussion

In this part, we present detailed discussions from two aspects to deeply recognize our proposed method.

(1) The illumination is a treasure not trash for the detector. Our proposed algorithm is established based on the fact that “the illumination is a treasure not trash for the detector”. Here we verify this fact from an experimental perspective. In Table 1, we investigate different training patterns for enhancer + detector. Here, we consider two representative enhancers (RUAS and SCI) and the classical SSD detector with Feature Pyramid Networks. The fine-tune detector and joint training methods cannot achieve good detection results. Fortunately, after introducing the illumination for the detector (i.e., RUAS^{++} and SCI^{++}), the detection accuracy for these two enhancers all realize a significant boost (see the red-bold texts in the right bottom corner of the third and last rows in Table 1). Deeply thinking, the performance improvement is benefited from the entire expression for scene information. In a word, the experiments can fully verify the necessity of introducing illumination for the detector. Compared with existing manners, we can conclude that “illumination is a treasure not trash for the detector”.

(2) Our T2 realizes the detection-oriented enhancement. In our designed algorithm, the scene decomposition module is constructed based on physical knowledge (i.e., Retinex theory) for low-light image enhancement. Although we do

Figure 4: Comparing decomposed components among different methods on a low-light image. Except the input, the bottom left and top right for each image are illumination and reflectance, respectively.

not define the loss functions related to visual quality in the training phase, this module still implicitly possesses the tendentiousness for enhancing image quality. To verify it, we show the decomposed components among different methods (most methods come from Table 1) on low-light scenarios, respectively.

Clearly, as seen in Figure 4, compared with the original enhanced outputs (i.e., RUAS and SCI), the other detection-oriented enhancers, even without visual quality-related constraints, all achieve enhancement effects. Notably, our fully designed T2 exhibits a certain image enhancement effect and is considered to possess friendly-detection features (as observed in the zoomed-in region). Moreover, RUAS^{++} and SCI^{++} perform better visual quality than RUAS^{+} and SCI^{+}, which also indicates that our T2 also realizes a mutual promotion for visual quality and detection accuracy (see Table 1).

Experimental Results

Implementation Details

Datasets and Metrics We conducted our experiments using the DARK FACE dataset, which consists of 6000 low-light images captured in real-world environments. The images have a resolution of 1080×720 and contain a variable number of faces, typically ranging from 1 to 20. The labeled faces exhibit a wide range of scales, varying from 1×2 to 335×296. For our experiments, we randomly selected 1000 images for testing, while the remaining images were used for training. To evaluate the performance of our approach, we employed the mean Average Precision (mAP) as the evaluation metric.

Parameters Setting For model training, we employed SGD with a momentum of 0.9 and weight decay of 0.0005. The batch size was set to 4, and the initial learning rate was 0.0005. During the training stage, a face anchor was labeled as a positive anchor if it had an Intersection over Union (IoU) of over 0.3 with the ground truth. Furthermore, we maintained a ratio of 3:1 between negative and positive anchors. To ensure consistency, we resized the images to 640×640 for training and 1500×1000 for testing. For effective bounding box selection, we applied non-maximum suppression using Jaccard overlap with a threshold of 0.3, retaining the top 750 high-confidence bounding boxes per image, as inspired by Neubeck and Van Gool's work (Neubeck and Van Gool 2006).

Compared Methods In order to provide a comprehensive evaluation of our detection results, we employed both qualitative and quantitative analyses. Our detection framework is built upon two state-of-the-art face detectors, namely Pyramidbox (Tang et al. 2018) and DSFD (Li et al. 2019). To enhance the quality of the input images, we utilized a variety of image enhancement methods including DeepUPE (Wang et al. 2019), DRBN (Yang et al. 2020a), EnGAN (Jiang et al. 2021), FIDE (Xu et al. 2020), KinD (Zhang, Zhang, and Guo 2019), RUAS (Liu et al. 2021c), ZeroDCE (Guo et al. 2020), and SCI (Ma et al. 2022b). In addition to these image enhancement methods, we also compared our results with two dedicated low-light face detection solutions, HLA (Wang, Yang, and Liu 2021) and REG (Liang et al. 2021).

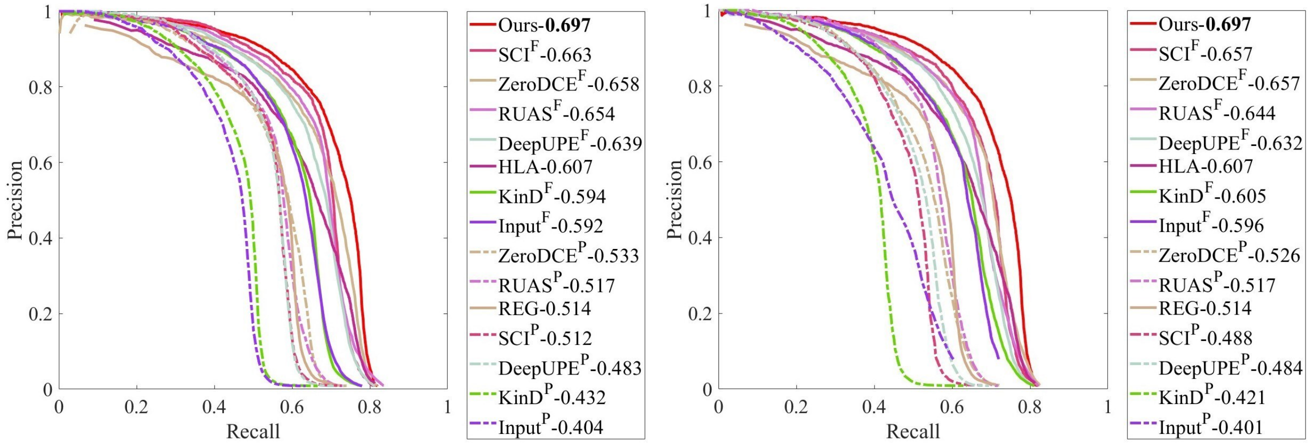
Evaluations on the DARK FACE Dataset

Quantitative Comparisons In-depth analysis was conducted by plotting Precision-Recall (PR) curves for two detectors, as illustrated in Figure 5. This comparison revealed two noteworthy conclusions.

Firstly, our proposed method exhibited significant advancements when compared to both the combination of advanced enhancer and detector, as well as mainstream low-light face detectors. To be more specific, our method achieved a remarkable 14.8% higher mean Average Precision (mAP) than HLA and an impressive 8.2% higher mAP than the best hybrid scheme, which involved the finetuning version of SCI and DSFD. Importantly, our proposed method consistently outperformed different detectors across various numerical metrics.

Secondly, these findings lend support to our initial motivation that solely relying on visual aesthetics may not necessarily lead to improved detection performance. Although the jointly fine-tuned methods demonstrated some enhancement in performance, there remains substantial room for further improvement.

Qualitative Comparisons We selected three representative scenarios to demonstrate our detection precision and visual effects straightforwardly, including extreme darkness, dense crowd, and low contrast. Figure 6 demonstrated four groups of visual and detection results comparisons on the DARK FACE dataset. We can observe that our methods obtain two obvious advantages compared with these competitors. Firstly, our method could effectively extract efficient semantic information from illumination for detection and re-



(a) Pyramidbox (b) DSFD

Figure 5: The PR curves of different state-of-the-art methods and our proposed approach on the DARK FACE dataset. $(\cdot)^{\mathrm{P}}$ and $(\cdot)^{\mathrm{F}}$ represent the version of using directly pre-trained detectors and fine-tune detectors, respectively.

Figure 6: More visual results of object detection on the DARK FACE dataset.

alize high accuracy (*e.g.,* the people in the distance in each group of examples). Secondly, benefiting from the guidance of follow-up detection, our method could decompose more suitable illumination to preserve the detection performance. In contrast, our method had less noise interference in detection.

Thus, by leveraging the decomposition and aggregation, the proposed method actually realized the representative feature extraction and utilization, which can simultaneously provide better scene understanding.

Evaluations on the ExDARK Dataset

In order to fully verify the performance of detection, we presented more results on the ExDark (Loh and Chan 2019) low-light object detection dataset. This dataset involves 12 categories. 737 images were randomly sampled for testing and the remaining 6626 low-light images were used for training and validation. For this dataset, we set the maximum epoch as 100, and the batch size as 32. We used Adam and the learning rate was initialized to 3e-5. We utilized the SSD detection model as the baseline in all methods for the ExDark dataset. We compared our method with the ”Enhancer+Detector” detection pattern, which considered low-light image enhancement as a pre-processing method (Ma et al. 2022c). Table 2 (score-threhold=0.5) reported the detection results on the specific class. It could be easily observed that our method was significantly superior to other methods. Furthermore, Figure 7 provided the visual comparison among different methods on the ExDark dataset, our method detected more correct objects and made fewer errors.

Algorithmic Analyses

Ablation Study We firstly provided five variants (denoted as “A$\sim$E”) under different settings to demonstrate the proposed mechanisms. Concrete numerical results were re-

Method	Bicycle	Boat	Bottle	Bus	Car	Cat	Chair	Cup	Dog	Motorbike	People	Table	mAP
Low-Light Input	61.87	54.42	41.77	85.55	62.14	57.28	45.81	39.30	57.58	63.68	54.29	51.88	56.30
DeepUPE	63.99	47.82	40.34	88.92	62.12	58.52	46.00	42.06	59.07	**64.28**	54.97	51.86	56.66
ZeroDCE	**68.04**	50.19	41.61	88.94	60.04	54.76	48.62	38.25	58.68	62.62	54.79	55.67	56.85
EnGAN	60.24	52.17	38.31	**89.82**	60.60	56.49	47.45	38.24	57.75	63.13	52.59	50.04	55.57
KinD	62.76	54.07	40.36	84.62	58.33	53.07	42.16	35.86	59.90	58.34	53.75	48.71	54.33
DRBN	63.54	53.14	36.35	86.38	58.83	51.66	39.85	33.73	55.50	59.87	52.12	50.23	53.44
RUAS	63.49	47.93	40.61	85.94	60.19	50.14	44.46	39.12	54.37	60.20	53.52	51.91	54.32
SCI	65.41	52.08	**42.31**	88.80	60.44	53.11	47.21	38.29	56.57	60.78	54.80	49.49	55.77
Ours	63.47	**57.87**	41.92	88.16	**67.09**	**64.24**	**51.93**	**42.85**	**60.40**	63.83	**58.69**	**56.48**	**59.74**

Table 2: Quantitative results of object detection on the ExDark dataset about the finetuned detector (SSD) on the enhanced results generated by all the compared methods. The best result is highlighted in bold.

Input DeepUPE DRBN EnGAN KinD RUAS SCI ZeroDCE Ours Ground Truth

Figure 7: More visual results of object detection on the ExDark dataset.

Model	**L**	**I**	**R**	**FPN**	**mAP (%)**
A	✓				61.13
B	✓			✓	66.42
C		✓			59.28
D			✓		66.94
E			✓	✓	67.68
T2		✓	✓	✓	**69.67**

Table 3: Ablation study among different settings.

ported in Table 3 (when both illumination and reflectance were used as input to the feature extractor, the FPN represented our semantic aggregation FPN). We can summarize three vital observations from this table. Firstly, the illumination layer also maintains enough information capacity, which almost obtains a comparable result compared with the version using low-light inputs **L**. Furthermore, reflection alone cannot obtain the best performance. Thus, it demonstrates the reasonability to extract the information from illumination. Secondly, FPN network plays an important role to improve the performance, which can significantly improve 8.6% and 0.74 higher mAP when the inputs are low-light images **L** and reflection **R** respectively. Therefore, we use the semantic aggregation FPN in T2. For the performance of using scene information including illumination and reflectance, accuracy is significantly improved.

Concluding Remarks

In this work, we proposed a new low-light object detector consisting of the scene decomposition and the semantic aggregation modules. We first analyzed the latent relationship between enhancer and detector to definitely point out the two key challenges. Then we constructed a scene decomposition module to present scene characteristics. A semantic aggregation composed of a weight-sharing feature extractor and multi-scale feature aggregator is proposed to integrate the scene information in the feature space. Finally, extensive experiments are performed to reveal our superiority.

Broader Impacts. How to effectively characterize scene information is an inherent appeal in designing our method. It is actually a general research focus for a variety of adverse conditions. Decomposing the scenes in the image space and aggregating the scenes in the feature space are crucial measures for our proposed T2. This way provides a new perspective to understand and handle the pattern of enhancer + detector, which will rekindle the enthusiasm to research the task of detection in adverse conditions.

Acknowledgments

This work is partially supported by the National Key R&D Program of China (No. 2022YFB1313500), the National Natural Science Foundation of China (Nos. U22B2052, 12222106 and 62302078), the LiaoNing Revitalization Talents Program (No. 2022RG04) and China Postdoctoral Science Foundation (No. 2023M730741).

References

Chen, C.; Chen, Q.; Xu, J.; and Koltun, V. 2018. Learning to see in the dark. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 3291–3300.

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*, 2553–2562.

Girshick, R. 2015. Fast r-cnn. In *Proceedings of the IEEE international conference on computer vision*, 1440–1448.

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*, 1780–1789.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.

Jiang, Y.; Gong, X.; Liu, D.; Cheng, Y.; Fang, C.; Shen, X.; Yang, J.; Zhou, P.; and Wang, Z. 2021. Enlightengan: Deep light enhancement without paired supervision. *IEEE Transactions on Image Processing*, 30: 2340–2349.

Jin, D.; Ma, L.; Liu, R.; and Fan, X. 2021. Bridging the gap between low-light scenes: Bilevel learning for fast adaptation. In *Proceedings of the 29th ACM International Conference on Multimedia*, 2401–2409.

Jobson, D. J.; Rahman, Z.-u.; and Woodell, G. A. 1997. A multiscale retinex for bridging the gap between color images and the human observation of scenes. *IEEE Transactions on Image processing*, 6(7): 965–976.

Land, E. H.; and McCann, J. J. 1971. Lightness and retinex theory. *Josa*, 61(1): 1–11.

Li, C.; Guo, J.; Porikli, F.; and Pang, Y. 2018. LightenNet: A convolutional neural network for weakly illuminated image enhancement. *Pattern recognition letters*, 104: 15–22.

Li, J.; Wang, Y.; Wang, C.; Tai, Y.; Qian, J.; Yang, J.; Wang, C.; Li, J.; and Huang, F. 2019. DSFD: dual shot face detector. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5060–5069.

Liang, J.; Wang, J.; Quan, Y.; Chen, T.; Liu, J.; Ling, H.; and Xu, Y. 2021. Recurrent exposure generation for low-light face detection. *IEEE Transactions on Multimedia*, 24: 1609–1621.

Lin, T.-Y.; Dollár, P.; Girshick, R.; He, K.; Hariharan, B.; and Belongie, S. 2017. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2117–2125.

Liu, J.; Fan, X.; Jiang, J.; Liu, R.; and Luo, Z. 2021a. Learning a deep multi-scale feature ensemble and an edge-attention guidance for image fusion. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(1): 105–119.

Liu, J.; Xu, D.; Yang, W.; Fan, M.; and Huang, H. 2021b. Benchmarking low-light image enhancement and beyond. *International Journal of Computer Vision*, 129: 1153–1184.

Liu, R.; Fan, X.; Zhu, M.; Hou, M.; and Luo, Z. 2020. Real-world underwater enhancement: Challenges, benchmarks, and solutions under natural light. *IEEE transactions on circuits and systems for video technology*, 30(12): 4861–4875.

Liu, R.; Lin, Z.; De la Torre, F.; and Su, Z. 2012. Fixed-rank representation for unsupervised visual learning. In *2012 ieee conference on computer vision and pattern recognition*, 598–605. IEEE.

Liu, R.; Liu, Z.; Mu, P.; Fan, X.; and Luo, Z. 2023a. Optimization-Inspired Learning with Architecture Augmentations and Control Mechanisms for Low-Level Vision. *IEEE Transactions on Image Processing*.

Liu, R.; Ma, L.; Ma, T.; Fan, X.; and Luo, Z. 2022a. Learning with nested scene modeling and cooperative architecture search for low-light vision. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Liu, R.; Ma, L.; Zhang, J.; Fan, X.; and Luo, Z. 2021c. Retinex-inspired unrolling with cooperative prior architecture search for low-light image enhancement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10561–10570.

Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S.; Fu, C.-Y.; and Berg, A. C. 2016. Ssd: Single shot multibox detector. In *European conference on computer vision*, 21–37. Springer.

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

Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; and Guo, B. 2021d. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, 10012–10022.

Liu, Z.; Liu, J.; Wu, G.; Ma, L.; Fan, X.; and Liu, R. 2023b. Bi-level Dynamic Learning for Jointly Multi-modality Image Fusion and Beyond. *arXiv preprint arXiv:2305.06720*.

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

Lore, K. G.; Akintayo, A.; and Sarkar, S. 2017. LLNet: A deep autoencoder approach to natural low-light image enhancement. *Pattern Recognition*, 61: 650–662.

Lv, F.; Li, Y.; and Lu, F. 2021. Attention guided low-light image enhancement with a large scale low-light simulation dataset. *International Journal of Computer Vision*, 129(7): 2175–2193.

Ma, L.; Jin, D.; An, N.; Liu, J.; Fan, X.; Luo, Z.; and Liu, R. 2023. Bilevel fast scene adaptation for low-light image enhancement. *International Journal of Computer Vision*, 1–19.

Ma, L.; Liu, R.; Wang, Y.; Fan, X.; and Luo, Z. 2022a. Low-light image enhancement via self-reinforced retinex projection model. *IEEE Transactions on Multimedia*.

Ma, L.; Liu, R.; Zhang, J.; Fan, X.; and Luo, Z. 2021. Learning deep context-sensitive decomposition for low-light image enhancement. *IEEE Transactions on Neural Networks and Learning Systems*, 33(10): 5666–5680.

Ma, L.; Ma, T.; Liu, R.; Fan, X.; and Luo, Z. 2022b. Toward Fast, Flexible, and Robust Low-Light Image Enhancement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5637–5646.

Ma, T.; Ma, L.; Fan, X.; Luo, Z.; and Liu, R. 2022c. 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.

Neubeck, A.; and Van Gool, L. 2006. Efficient non-maximum suppression. In *18th International Conference on Pattern Recognition (ICPR'06)*, volume 3, 850–855. IEEE.

Piao, Y.; Ji, W.; Li, J.; Zhang, M.; and Lu, H. 2019. Depth-induced multi-scale recurrent attention network for saliency detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, 7254–7263.

Piao, Y.; Rong, Z.; Zhang, M.; Ren, W.; and Lu, H. 2020. A2dele: Adaptive and attentive depth distiller for efficient RGB-D salient object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 9060–9069.

Tang, H.; Li, Z.; Peng, Z.; and Tang, J. 2020. Blockmix: meta regularization and self-calibrated inference for metric-based meta-learning. In *Proceedings of the 28th ACM international conference on multimedia*, 610–618.

Tang, H.; Yuan, C.; Li, Z.; and Tang, J. 2022. Learning attention-guided pyramidal features for few-shot fine-grained recognition. *Pattern Recognition*, 130: 108792.

Tang, X.; Du, D. K.; He, Z.; and Liu, J. 2018. Pyramidbox: A context-assisted single shot face detector. In *Proceedings of the European conference on computer vision (ECCV)*, 797–813.

Wang, R.; Zhang, Q.; Fu, C.-W.; Shen, X.; Zheng, W.-S.; and Jia, J. 2019. Underexposed photo enhancement using deep illumination estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6849–6857.

Wang, W.; Wang, X.; Yang, W.; and Liu, J. 2022. Unsupervised face detection in the dark. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(1): 1250–1266.

Wang, W.; Yang, W.; and Liu, J. 2021. Hla-face: Joint high-low adaptation for low light face detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 16195–16204.

Wei, C.; Wang, W.; Yang, W.; and Liu, J. 2018. Deep retinex decomposition for low-light enhancement. *arXiv preprint arXiv:1808.04560*.

Wu, J.; Lin, Z.; and Zha, H. 2019. Essential tensor learning for multi-view spectral clustering. *IEEE Transactions on Image Processing*, 28(12): 5910–5922.

Xu, K.; Yang, X.; Yin, B.; and Lau, R. W. 2020. Learning to restore low-light images via decomposition-and-enhancement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2281–2290.

Xue, X.; He, J.; Ma, L.; Wang, Y.; Fan, X.; and Liu, R. 2022. Best of Both Worlds: See and Understand Clearly in the Dark. In *Proceedings of the 30th ACM International Conference on Multimedia*, 2154–2162.

Yang, W.; Wang, S.; Fang, Y.; Wang, Y.; and Liu, J. 2020a. From fidelity to perceptual quality: A semi-supervised approach for low-light image enhancement. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 3063–3072.

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

Yuan, Y.; Yang, W.; Ren, W.; Liu, J.; Scheirer, W. J.; and Wang, Z. 2019. UG2+Track 2: A Collective Benchmark Effort for Evaluating and Advancing Image Understanding in Poor Visibility Environments. *arXiv preprint arXiv:1904.04474*.

Zhang, M.; Ren, W.; Piao, Y.; Rong, Z.; and Lu, H. 2020. Select, supplement and focus for RGB-D saliency detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 3472–3481.

Zhang, S.; Zhu, X.; Lei, Z.; Shi, H.; Wang, X.; and Li, S. Z. 2017. S3fd: Single shot scale-invariant face detector. In *Proceedings of the IEEE international conference on computer vision*, 192–201.

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