You Only Look Around: Learning Illumination Invariant Feature for Low-light Object Detection

Mingbo Hong*
Megvii Technology, Beijing, China
mingbohong97@gmail.com

Shen Cheng*
Megvii Technology, Beijing, China
chengshen@megvii.com

Haibin Huang
Kuaishou Technology
jackiehuanghaibin@gmail.com

Haoqiang Fan
Megvii Technology, Beijing, China
fhq@megvii.com

Shuaicheng Liu†
University of Electronic Science and Technology of China
liushuaicheng@uestc.edu.cn

Abstract

In this paper, we introduce YOLA, a novel framework for object detection in low-light scenarios. Unlike previous works, we propose to tackle this challenging problem from the perspective of feature learning. Specifically, we propose to learn illumination-invariant features through the Lambertian image formation model. We observe that, under the Lambertian assumption, it is feasible to approximate illumination-invariant feature maps by exploiting the interrelationships between neighboring color channels and spatially adjacent pixels. By incorporating additional constraints, these relationships can be characterized in the form of convolutional kernels, which can be trained in a detection-driven manner within a network. Towards this end, we introduce a novel module dedicated to the extraction of illumination-invariant features from low-light images, which can be easily integrated into existing object detection frameworks. Our empirical findings reveal significant improvements in low-light object detection tasks, as well as promising results in both well-lit and over-lit scenarios. Code is available at https://github.com/MingboHong/YOLA.

1 Introduction

In the field of computer vision, object detection stands as a cornerstone, driving advancements in numerous applications ranging from autonomous vehicles to security surveillance [26, 51, 20]. The ability to accurately identify and locate objects in digital imagery has seen remarkable progress, largely due to the advent of deep learning techniques [16, 15, 40]. However, despite these advancements, object detection in low-light conditions remains a significant challenge. Low-light environments lead to poor image quality, reduced visibility, and increased misdetections in night-time surveillance and twilight driving [48, 32].

Traditional methods in tackling low-light object detection have predominantly leaned towards image enhancement techniques [17, 24, 53, 34]. While these methods have demonstrated effectiveness in

*Equal contribution.
†Corresponding Author.

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

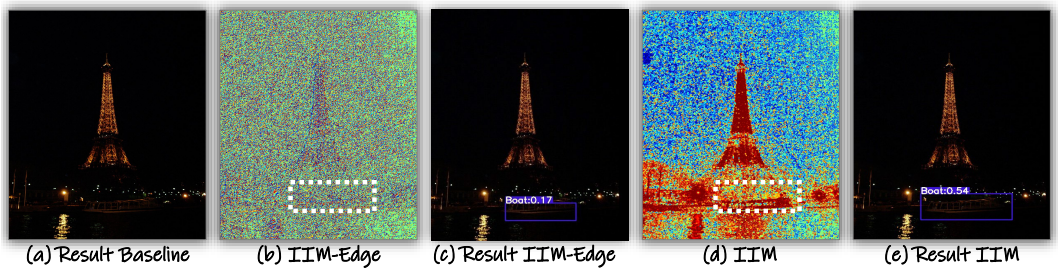


Figure 1: (a): The base detector failed to recognize objects. (b, c) However, when IIM is employed with a simple edge feature, the object is identified. (d, e) Furthermore, the full IIM utilizes a task-driven learnable kernel to extract illumination-invariant features that are richer and more suitable for the detection task than simple edge features.

improving visual aesthetics and perceptual quality, they often do not directly translate to improved object detection performance. This discrepancy arises because these enhancement techniques are typically optimized for human visual perception, which does not always correlate with the requirements for effective and accurate object detection by machine learning models.

In addition to image enhancement strategies, another research direction involves fine-tuning pre-trained models for low-light conditions. Typically, detectors are initially trained on extensive datasets of well-lit images, such as those from Pascal VOC [11] and Microsoft COCO [28], and subsequently fine-tuned on smaller, low-light datasets [48, 32]. To enhance the utilization of cross-domain information, the MAET framework [7] was developed to learn intrinsic visual structure features by separating object features from those caused by degradation in image quality. Similarly, methods [31, 25] aim to restore the normal appearances of corrupted images during detector training. However, these techniques often depend heavily on synthetic datasets, which could limit their real-world applicability.

Recent methods in low-light object detection, such as those in [36, 49], use Laplacian pyramids [2] for multi-scale edge extraction and image enhancement. FeatEnHancer [18] further leverages hierarchical features for improved low-light vision. However, these task-specific, loss-driven approaches often grapple with a larger solution space due to varying illumination effects.

In this study, we introduce a novel approach that explicitly leverages illumination-invariant features, utilizing the principles of the Lambertian image formation model [42]. Under the Lambertian assumption, we can express the pixel values in each channel as a discrete combination of three key components: the surface normal, the light direction (both of which are solely related to the pixel's position), the spectral power distribution, and the intrinsic properties of the pixel itself. The illumination-invariant feature can be learned by eliminating the position-related term and spectral power-related term [14]. We introduce this concept of extracting illumination-invariant features into low-light detection tasks and demonstrate that incorporating this feature yields significant performance improvements in low-light detection tasks. We further improve this illumination-invariant feature using task-driven kernels. Our key observation is that by imposing a zero-mean constraint on these kernels, the feature can simultaneously discover richer downstream task-specific patterns and maintain illumination invariance, thereby improving performance.

Towards this end, we propose the Illumination-Invariant Module (IIM), a versatile and adaptive component designed to integrate the information gleaned from these specialized kernels with standard RGB images. The IIM can be seamlessly integrated with a variety of existing object detection frameworks, enhancing their capability to perform accurately in low-light environments, whether through naive edge features or diverse illumination-invariant characteristics, as shown in Fig 1. We further conduct experiments on the ExDark and UG^2+DARK FACE datasets to evaluate our method. Our experimental results demonstrate that the integration of the IIM significantly enhances the detection accuracy of existing methods, leading to substantial improvements in low-light object detection. To summarize, our contributions are as follows:

- We introduce YOLA, a novel framework for object detection in low-light conditions by leveraging illumination-invariant features.

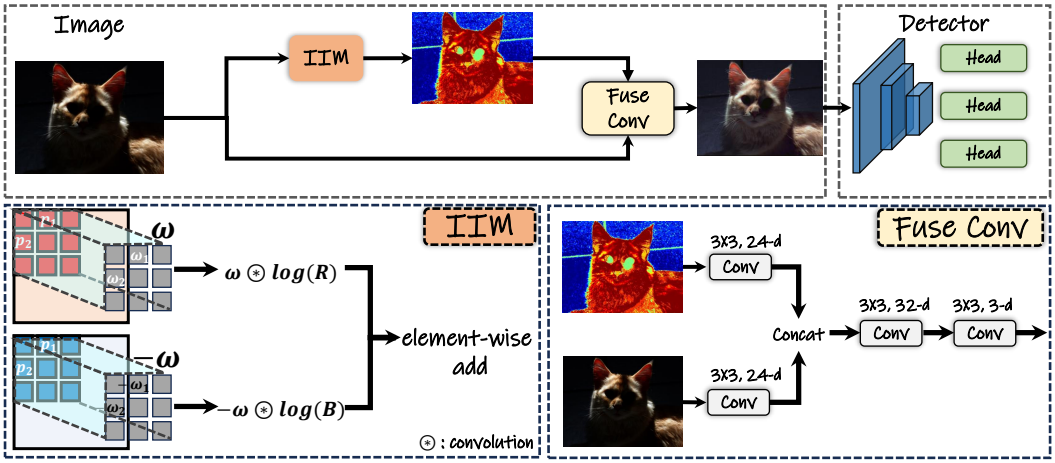


Figure 2: The overall pipeline of YOLA.YOLA extracts illumination-invariant features via IIM and integrates them with original images by leveraging a fuse convolution block for the subsequent detector.

- We design a novel Illumination-Invariant Module to extract illumination-invariant features without requiring additional paired datasets, and can be seamlessly integrated into existing object detection methods.
- We provide an in-depth analysis of the extracted illumination-invariant paradigm and propose a learning illumination-invariant paradigm.
- Our experiments show YOLA can significantly improve the detection accuracy of existing methods when dealing with low-light images.

2 Related work

2.1 General object detection

Current modern object detection methods can be classified as anchor-based and anchor-free. The anchor-based detectors are derived from the sliding-window paradigm, where the dense anchor can be viewed as the sliding-window arranged in spatial space. Subsequently, the anchors are assigned as positive or negative samples based on the matching strategy (i.e., Intersection-over Union (IoU) [16], Top-K [52, 50]). Common anchor-based methods include R-CNN [16, 15, 40], SSD [30], YOLOv2 [38], and RetinaNet [27], among others. In contrast, the anchor-free detectors liberate the handcraft anchor hyper-parameter setting, enhancing their potential in terms of generalization capability. Prominent methods in anchor-free include YOLOv1 [37], FCOS [44], and DETR [3]. Despite the remarkable achievements of both anchor-based and anchor-free detectors in general object detection, they exhibit unsatisfactory performance under low-light conditions.

2.2 Low-light object detection

Object detection in low-light conditions remains a significant challenge. One common line of research involves leveraging image enhancement techniques, such as KIND [53], SMG [46], NeRCo [47], and others [17, 24, 22, 23] to directly improve the quality of the low-light image. The enhanced images are then deployed in the subsequent training and testing stages of detection. However, the objective of image enhancement is inherently different from that of object detection, making this strategy suboptimal. To address this, some researchers [21, 6] explore integrating image enhancement with object detection during the training process. Nevertheless, the task of balancing hyperparameters to equilibrate visual quality and detection performance remains intricate. Recently, Sun et al. [43] proposed a targeted adversarial attack paradigm aimed at restoring degraded images to ones that are more favorable for object detection. MAET [7] trained on a low-light synthetic dataset, obtaining the pre-trained model endowed intrinsic structure decomposition ability for downstream lowlight object detection. Further, IA-YOLO [31] and GDIP [25] elaborately design the differentiable image processing module to enhance image adaptively for adverse weather object detection. Note that the

aforementioned methods either require a dedicated low-light enhancement dataset or rely heavily on synthetic datasets in training. To mitigate the limitations, a set of methodologies [36, 49, 18] utilize multi-scale hierarchical features and are driven purely by task-specific loss to improve low-light vision. Unlike those methods, we introduce illumination-invariant features to alleviate the effect of illumination on low-light object detection, without requiring additional low-light enhancement datasets or synthetic datasets.

2.3 Illumination invariant representation

Adverse illumination typically degrades the performance on downstream tasks, prompting researchers to explore illumination-invariant techniques to mitigate this impact. For high-level tasks, Wang et al. [45] proposed an illumination normalization method for Face Recognition. Alshammari et al. [1] use illumination-invariant image representation to improve automotive scene understanding and segmentation. Lu et al. [33] convert RGB images to illumination-invariant chromaticity space, preparing for the following feature extraction to achieve traffic object detection in various illumination conditions. For low-level tasks, several physics-based invariants, such as Colour Ratios [13] (CR) and Cross Colour Ratios [14] (CCR), are employed to decompose the illumination for intrinsic image decomposition [10, 9, 8]. However, these methods leverage illumination-invariant representations derived from the fixed formulations, which may not adequately capture the diverse and complex illumination scenarios that are specific to downstream applications. In contrast, our method enables the adaptive learning of illumination-invariant features in an end-to-end manner, thereby enhancing compatibility with downstream tasks.

3 Method

In this section, we formally introduce YOLA, a novel method for low-light object detection. As illustrated in Fig.2, the key component of YOLA is the Illumination Invariant Module (IIM) focusing on feature learning to derive downstream task-specific illumination-invariant features. These features can be integrated with existing detection modules, enhancing their capability in low-light conditions. Next, we will introduce the derivation of illumination-invariant features and provide a detailed description of IIM's specific implementation.

3.1 Illumination invariant feature

Notation: Let I represents an image in the standard RGB domain, and let $C \in [R, G, B]$ represent the image in the red, green, or blue channel. We define the value in channel C of a pixel p_i as C_{p_i}, where $i \in I$ is the pixel index.

Lambertian assumption: According to body reflection term of the dichromatic reflection model, the value of C_{p_i} can be expressed in the discreet form as follows:

$$C_{p_i} = m(\vec{n}_{p_i}, \vec{l}_{p_i}) e^{C_{p_i}}(\lambda) \rho^{C_{p_i}}(\lambda), \tag{1}$$

Here, $\vec{n}_{p_i}, \vec{l}_{p_i}$ represents surface normal and light direction respectively, and m denotes the interaction function between them. The term $e^{C_{p_i}}$ represents the spectral power distribution of the illuminant at point p_i in color channel C , and $\rho^{C_{p_i}}$ represents the intrinsic property (reflectance) of the object at point p_i in color channel C.

It becomes apparent that the term m is determined solely by the positional component, with no impact from the color channels. This observation leads to the strategy of calculating the difference between values of different color channels at the same spatial positions to effectively eliminate the influence of m. To eliminate the term e, we can utilize the assumption that illumination is approximately uniform across adjacent pixels. Consequently, by computing the difference between values of neighboring pixels, we can further further eliminate the influence of m.

Cross color ratio: Taking into consideration two adjacent pixels, denoted as p_1 and p_2, along with the red (R) and blue (B) channels, we can determine the ratio M_{rb} between the red and blue channels through the following computational procedure:

$$M_{rb} = \frac{R_{p_1} B_{p_2}}{R_{p_2} B_{p_1}}. \tag{2}$$

Taking the logarithm of M_{rb} and substituting the pixel values with Eq. 1, we get:

$$\begin{aligned} log(M_{rb}) = log(m(\vec{n_{p_1}}, \vec{l_{p_1}})) - log(m(\vec{n_{p_1}}, \vec{l_{p_1}})) \\ + log(e^{R_{p_1}}(\lambda)) - log(e^{R_{p_2}}(\lambda)) \\ + log(\rho^{R_{p_1}}(\lambda)) - log(\rho^{R_{p_2}}(\lambda)) \\ + log(m(\vec{n_{p_2}}, \vec{l_{p_2}})) - log(m(\vec{n_{p_2}}, \vec{l_{p_2}})) \\ + log(e^{B_{p_2}}(\lambda)) - log(e^{B_{p_1}}(\lambda)) \\ + log(\rho^{B_{p_2}}(\lambda)) - log(\rho^{B_{p_1}}(\lambda)). \end{aligned} \tag{3}$$

With the illumination assumption that $e^{C_{p_1}} \approx e^{C_{p_2}}$, the above equation can be further simplified into an illumination-invariant form:

$$\begin{aligned} log(M_{rb}) = log(\rho^{R_{p_1}}(\lambda)) - log(\rho^{R_{p_2}}(\lambda)) \\ + log(\rho^{B_{p_2}}(\lambda)) - log(\rho^{B_{p_1}}(\lambda)) \end{aligned} \tag{4}$$

By observing the elimination in Eq. 4, we can find that subtraction within the **same channel** eliminates the illumination term (implemented by zero-mean constraint), while **cross-channel** subtraction removes the surface normal and light direction terms, which motivates us to design the learning illumination-invariant paradigms.

In this case, we can use a convolution operation to extract features, as shown in Fig. 2. The extracted features are processed and fused by the IIM before being sent to the detector. When using fixed weights of adjacent pixels with a subtraction value of 1 or -1, we refer to it as IIM-Edge. Next, we will provide a detailed introduction to the IIM.

3.2 Illumination invariant module

While Eq. 4 offers a straightforward and effective method for calculating Illumination Invariant features, its rigidity presents certain limitations. Specifically, the fixed nature of this equation may not adequately capture the diverse and complex variations in illumination that are specific to downstream tasks across different scenarios. To address this, we have evolved the equation into a more adaptable form using convolutional operations. Instead of relying on a single kernel, our approach involves learning a set of convolutional kernels. This strategy not only enhances the robustness of the Illumination Invariant feature extraction but also improves its efficiency. To this end, we propose Illumination Invariant Module comprising two main components, including learnable kernels and a zero-mean constraint. Note that Illumination Invariant Module yield features are inherently illumination invariant at initialization. Subsequent kernel learning is geared towards producing task-specific illumination invariant features for downstream tasks.

Learnable kernel. The goal is to transform the fixed illumination-invariant feature into a learnable form. Specifically, we aim to learn a set of convolutional kernels $\mathcal{W}_1, \mathcal{W}_2, \cdots \mathcal{W}_n^{\in k \times k}$, where n represents the number of kernels and k denotes the kernel size. Here, we extend the fixed feature into a more versatile and generalized form. Let p_i and w_i represent a group pixel position and its corresponding weight within a kernel $\mathcal{W}_n$, where $i = 0, 1, \cdots k^2$. These parameters enable us to evolve the Cross Color Ratio (CCR) into an adaptable form, enhancing its capability to effectively handle varying illumination conditions. Note that w_i is trainable, rendering the positive or negative polarity inconsequential.

$$M_{rb} = \prod_{i=1}^{k^2-1} \left(\frac{R_{p_i}}{B_{p_i}}\right)^{w_i} \left(\frac{B_{p_{i+1}}}{R_{p_{i+1}}}\right)^{w_{i+1}} = \prod_{i=1}^{k^2} \left(\frac{R_{p_i}}{B_{p_i}}\right)^{w_i} \tag{5}$$

To make the extended form still satisfy Illumination Invariant, the logarithm of M_{rb} should satisfy the following constraints:

$$\begin{cases} \sum_i^{k^2} w_i log(e^{R_{p_i}}(\lambda)) = 0 \\ \sum_i^{k^2} w_i log(e^{B_{p_i}}(\lambda)) = 0 \end{cases} \tag{6}$$

Figure 3: Qualitative comparisons of TOOD detector on both ExDark and UG^2+DARK FACE dataset, where the top 2 rows visualize the detection results from ExDark, and the bottom 2 rows show the results from UG^2+DARK FACE. The images are being replaced with enhanced images generated by LLIE or low-light object methods. Red dash boxes highlight the inconspicuous cases. Zoom in red dash boxes for the best view.

If the above equation holds true, the e term and the m term are eliminated. The final feature can be expressed in a generalized form:

$$log(M_{rb}) = \sum_{i}^{k^2} w_i log(\rho^{R_{p_i}}(\lambda)) - \sum_{i}^{k^2} w_i log(\rho^{B_{p_i}}(\lambda)) \tag{7}$$

Similarly, we can obtain $log(M_{rg})$ and $log(M_{gb})$ to form $f_{\mathcal{W}_i}(I)$.

The resulting features obtained by applying the kernel $\mathcal{W}_i$ to the image I denoted as $f_{\mathcal{W}_i}(I)$, can be expressed as:

$$f_{\mathcal{W}_i}(I) = \begin{bmatrix} \mathcal{W}_i \circledast log(R) + (-\mathcal{W}_i) \circledast log(B) \\ \mathcal{W}_i \circledast log(R) + (-\mathcal{W}_i) \circledast log(G) \\ \mathcal{W}_i \circledast log(G) + (-\mathcal{W}_i) \circledast log(B) \end{bmatrix} \tag{8}$$

where $\circledast$ denotes the convolution.

Zero mean constraint: Drawing from Eq. 6 and the approximation $e^{R_{p_{1i}}} \approx e^{B_{p_i}}$, in the context of convolutional kernels, we simply ensure that the mean of $\mathcal{W}_n^{\in k \times k}$ to be 0, as depicted by:

$$\overline{\mathcal{W}_n} = \frac{1}{k^2} \sum_{i=1}^{k^2} w_i = 0 \tag{9}$$

This constraint is enforced by substituting the mean value from the kernel $\mathcal{W}_n = \mathcal{W}_n - \overline{\mathcal{W}_n}$.

Methods	YOLOv3		TOOD	
	recall	mAP_{50}	recall	mAP_{50}
Baseline	84.6	71.0	91.9	72.5
KIND [53]	83.3	69.4	92.1	72.6
SMG [46]	82.3	68.5	91.8	71.5
NeRCo [47]	83.4	68.5	91.8	71.8
DENet [36]	84.2	71.3	92.6	73.5
GDIP [53]	84.8	72.4	92.2	72.8
IAT [53]	85.0	72.6	92.9	73.0
MAET [7]	85.1	72.5	92.5	74.3
YOLA-Naive	84.8	71.6	91.8	71.6
YOLA	**86.1**	**72.7**	**93.8**	**75.2**

Table 1: Quantitative comparisons of the ExDark dataset based on YOLOv3 and TOOD detectors.

Methods	YOLOv3		TOOD	
	recall	mAP_{50}	recall	mAP_{50}
Baseline	77.9	60.0	81.5	62.1
KIND [53]	76.0	58.4	82.4	63.8
SMG [46]	69.3	48.9	77.1	55.8
NeRCo [47]	68.9	49.1	76.8	55.6
DENet [36]	77.7	60.0	84.1	66.2
GDIP [53]	77.8	60.4	82.1	62.9
IAT [53]	77.6	59.8	82.1	62.0
MAET [7]	77.9	59.9	83.6	64.8
YOLA-Naive	76.6	59.2	82.8	64.6
YOLA	**79.1**	**61.5**	**84.9**	**67.4**

Table 2: Quantitative comparisons of the UG^2+DARK FACE dataset based on YOLOv3 and TOOD detectors.

4 Experiments

4.1 Implementation details

We evaluate the proposed method using the popular anchor-based detector YOLOv3 [39] and the anchor-free detector TOOD [12]. Both detectors are initially pre-trained on the COCO dataset and subsequently fine-tuned on the target datasets utilizing the SGD [41] optimizer with an initial learning rate of $1e$-3. Specifically, we resize the ExDark dataset images to 608×608 and train both detectors for 24 epochs, reducing the learning rate by a factor of 10 at epochs 18 and 23. For the UG^2+DARK FACE dataset, we resize images to 1500×1000 for TOOD and maintain the 608×608 resolution for YOLOv3 to be consistent with MAET. YOLOv3 is trained for 20 epochs, with the learning rate decreased by a factor of 10 at 14 and 18 epochs. TOOD are trained for 12 epochs, with the learning rate decreased by a factor of 10 at 8 and 11 epochs. Additionally, we implement a straightforward illumination-invariant model, denoted as **YOLA-Naive**, by removing the IIM and ensuring various illumination features are consistently imposed by an MSE loss. We implement YOLA using the MMDetection toolbox [4].

4.2 Dataset

We evaluate our proposed method on both real-world scenarios datasets: exclusively dark [32] (ExDark) and UG^2+DARK FACE [48]. ExDark dataset contains 7363 images ranging from low-light environments to twilight, including 12 categories, 3,000 images for training, 1,800 images for validation, and 2,563 images for testing. We calculate the overall mean average precision (mAP_{50}) and mean recall at the IoU threshold of 0.5 as the evaluation metric. UG^2+DARK FACE dataset contains 6,000 labeled face bounding box images, where 5,400 images are allocated for training and 600 images are reserved for testing, and calculating the corresponding recall and mAP_{50} as evaluation metrics. Additionally, we also evaluate the generalization ability of our method on the COCO 2017 [28] dataset.

4.3 Low-light object detection

Table 1 presents the quantitative results of YOLOv3 and TOOD detectors on the ExDark dataset, respectively. We report the low-light image enhancement (LLIE) methods, including KIND, SMG, and NeRCo, along with the state-of-the-art low-light object detection methods, DENet, and MAET. Compared to the low-light object detection methods, the LLIE methods fail to achieve satisfactory performance due to inconsistency between human visual and machine perception. The enhancement methodologies prioritize human preferences. However, it is important to note that optimizing for enhanced visual appeal may not align with optimized object detection performance. Despite being the current state-of-the-art in image enhancement techniques, SMG and NeRCo exhibit worse performance compared to KIND when evaluated in the context of object detection tasks. In contrast, end-to-end approaches such as DENet and MAET, which account for machine perception, generally yield superior results in object detection compared to the LLIE methods. Nevertheless, our method remains simple and effective when compared to similar approaches in the same category. Moreover, compared to YOLA-Naive, YOLA exhibits superior performance because its extracted features inherently possess illumination invariance, implying a smaller solution space compared to YOLA-Naive. Specifically, our method achieves the best performance on both anchor-based YOLOv3

Dataset	IIM	IIM-Edge	$\mathcal{Z}_{mean}$	mAP_{50}
Exdark				72.5
		✓		73.8
	✓			74.7
	✓		✓	**75.2**
DarkFace				62.1
		✓		64.5
	✓			66.9
	✓		✓	**67.4**

Table 3: The effectiveness of IIM, IIM-Edge and the zero mean constraint $\mathcal{Z}_{mean}$ based on TOOD. The blank line denotes the baseline.

Dataset	Method	AP_{50}	AP_{75}	mAP
well-lit	TOOD	59.0	45.3	41.7
	+ YOLA	**59.4**	**46.0**	**42.3**
over-light	TOOD	57.4	43.8	40.5
	+ YOLA	**58.3**	**44.6**	**41.2**

Table 4: Ablation study for YOLA on COCO `2017val`.

Method	Kind	SMG	NeRco	DENet	MAET	Ours
Size(M)	8.21	17.90	23.30	0.04	40	**0.008**

Table 5: Model size of different methods.

and anchor-free detectors TOOD, surpassing the baseline by significant gains of 1.7 and 2.5 mAP, indicative of its superiority and effectiveness. Meanwhile, compared with most LLIE and lowlight object detection techniques, the number of parameters in our YOLA (**0.008M**) is significantly lower, as presented in Table 5. This highlights the potential for our method to be deployed in lightweight practical applications. For a more detailed quantitative comparison, please refer to our appendix.

4.4 Low-light face detection

We have shown the results on the ExDark dataset. Next, we showcase the results on a dataset that includes small-sized objects. Table 2 presents the quantitative results of the detector YOLOv3 and TOOD on UG^2+DARK FACE dataset. Significantly, it is worth noting that most LLIE methods integrated into the YOLOv3 detector fail to achieve satisfactory results. This implies that the utilization of enhancement-based approaches can impair the details of small faces, hindering the learning of useful representations in such images. On the other hand, methods considering the object detection task demonstrate better performance, where YOLA increases the 1.5 mAP, demonstrating its superior performance and generalization capability. For the recently advanced detector TOOD, our method still outperforms these LLIE and low-light object detection methods, achieving a remarkable mAP of 67.4. This underscores YOLA's superior generalization capabilities in improving the performance of both anchor-based and anchor-free detection paradigms.

4.5 Quantitative results

The top 2 rows of Figure 3 show the qualitative results from the ExDark dataset using the TOOD detector, where existing methods exhibit missed detections, highlighted by the red dashed boxes. In contrast, YOLA excels in detecting these challenging cases, demonstrating its superior performance in complex scenarios. The bottom 2 rows exhibit the qualitative results of the UG^2+DARK FACE dataset using the TOOD detector. These faces are typically tiny under low-light conditions, making it difficult for most methods to achieve comprehensive results.

Although our method does not explicitly constrain image brightness, the enhanced images tend to display increased brightness in the final results. The visual results shown in the figures may appear slightly grayish due to the absence of value range constraints on the enhanced images. For image display, we conducted channel-wise normalization.

4.6 Ablation studies

4.6.1 Illumination invariant module

We evaluate the effectiveness of the IIM in detectors TOOD, as presented in Table 3, respectively. The 1st and 5th rows of Table 3 show the baseline detectors evaluated on ExDark and UG^2+DARK FACE dataset. By incorporating the IIM to introduce illumination-invariant features, the detector yields considerable performance gains (2.3 and 4.8 mAP for ExDark and UG^2+DARK FACE, respectively).

4.6.2 Zero mean constraint

By imposing a zero mean constraint on the convolutional kernels, the subtraction formed by the kernels can factor out the illumination items. To evaluate the impact of this constraint, we exclude it from IIM, and the results are shown in Table 3. It is evident that the removal of this constraint leads

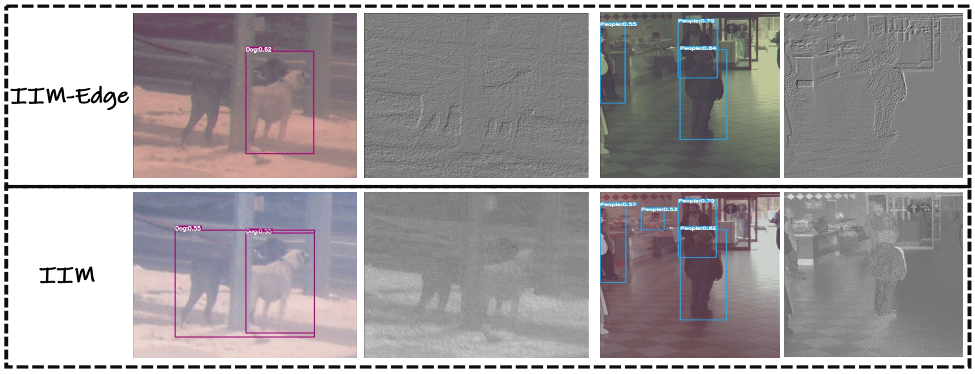


Figure 4: Visualization of the features (columns 2 and 4) generated by IIM-Edge and IIM(kernels are normalized for better visibility, we average the features across the channel dimensions and then conduct spatial normalization), along with detection results (columns 1 and 3). Best viewed by zooming in.

to a decline in performance, with reductions of 0.3 and 0.5 mAP for TOOD. These results indicate that utilizing the zero mean constraint to mitigate the effects of illumination is beneficial to low-light object detection.

4.6.3 Learnable kernel

The IIM is formed with the learnable kernels, encouraging the illumination-invariant features that are adaptively learned in an end-to-end fashion. In this experiment, we evaluate the fixed kernels (as specified in Eq. 4, also referred to as IIM-Edge), the results of which are shown in Table 3. It outperforms the baseline by 1.3 mAP on ExDark and 2.4 mAP on UG^2+DARK FACE, which demonstrates that the incorporation of illumination-invariant features is beneficial for low-light object detection. Subsequently, we substitute the fixed kernels with the learnable kernels, yielding further gains of 1.4 mAP on ExDark and 2.9 mAP on UG^2+DARK FACE. These results clearly prove the effectiveness of learnable kernels. In addition, we also impose a consistency loss for IIM's output feature to stabilize the kernel learning to prevent trivial solutions within the kernel, mitigating the impact of uneven lighting. (please refer to the appendix A for details).

Visualization: Illumination-invariant features exhibit considerable diversity, but the diversity captured by fixed kernels is limited. We visualize and compare the fixed kernel and learnable kernel as shown in Fig. 4. The features yielded by fixed kernels appear relatively uniform, primarily consisting of simple edge features. In contrast, learnable kernels extract more diverse patterns, resulting in visually richer and more informative representations.

4.7 Generalization

In this section, we broaden the application of the YOLA to the general object detection dataset COCO 2017, investigating the YOLA's generalization capability beyond low-light object detection. The metrics mAP (average for IoU [0.5:0.05:0.95]), AP_{50}, and AP_{75} are adopted to evaluate performance on COCO `2017val` (also called `minival`) as presented in Table 4. Specifically, we trained 12 epochs with 8 GPUs and a mini-batch of 1 per GPU in an initial learning rate of $1e$-2 by the SGD optimizer on both well-lit and over-lit (generated by brightening the origin image) scenarios. By observing Table 4, we can see that detectors integrated with YOLA in both scenarios exhibit notable improvements in performance.

5 Conclusion

In this work, we have revisited the complex challenge of object detection in low-light conditions and demonstrated the effectiveness of illumination-invariant features in improving detection accuracy in such environments. Our key innovation, the Illumination-Invariant Module (IIM), harnesses these features to great effect. By integrating a zero-mean constraint within the framework, we have effectively learned a diverse set of kernels. These kernels are adept at extracting illumination-invariant

features, significantly enhancing detection precision. We believe that our developed IIM module can be instrumental in advancing low-light object detection tasks in future applications.

Acknowledgement: This work was supported in part by National Natural Science Foundation of China (NSFC) under grant No.62372091 and Natural Science Foundation of Sichuan Province under grant Nos. 2023NSFSC0462 and 2023NSFSC1972.

References

[1] Naif Alshammari, Samet Akcay, and Toby P Breckon. On the impact of illumination-invariant image pre-transformation for contemporary automotive semantic scene understanding. In *2018 IEEE Intelligent Vehicles Symposium (IV)*, pages 1027–1032. IEEE, 2018.

[2] Peter J Burt and Edward H Adelson. The laplacian pyramid as a compact image code. In *Readings in computer vision*, pages 671–679. Elsevier, 1987.

[3] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *ECCV*, pages 213–229. Springer, 2020.

[4] Kai Chen, Jiaqi Wang, Jiangmiao Pang, Yuhang Cao, Yu Xiong, Xiaoxiao Li, Shuyang Sun, Wansen Feng, Ziwei Liu, Jiarui Xu, et al. Mmdetection: Open mmlab detection toolbox and benchmark. *arXiv preprint arXiv:1906.07155*, 2019.

[5] Linwei Chen, Ying Fu, Kaixuan Wei, Dezhi Zheng, and Felix Heide. Instance segmentation in the dark. *IJCV*, pages 1–21, 2023.

[6] 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 *BMVC*, page 238, 2022.

[7] 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 *ICCV*, pages 2553–2562, 2021.

[8] Partha Das, Maxime Gevers, Sezer Karaoglu, and Theo Gevers. Idtransformer: Transformer for intrinsic image decomposition. In *ICCV*, pages 816–825, 2023.

[9] Partha Das, Sezer Karaoglu, and Theo Gevers. Intrinsic image decomposition using physics-based cues and cnns. 223:103538, 2022.

[10] Partha Das, Sezer Karaoglu, and Theo Gevers. Pie-net: Photometric invariant edge guided network for intrinsic image decomposition. In *CVPR*, pages 19790–19799, 2022.

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

[12] Chengjian Feng, Yujie Zhong, Yu Gao, Matthew R Scott, and Weilin Huang. Tood: Task-aligned one-stage object detection. In *ICCV*, pages 3490–3499. IEEE Computer Society, 2021.

[13] Graham David Finlayson. Colour object recognition. *Simon Fraser University*, 1992.

[14] Theo Gevers and Arnold WM Smeulders. Color-based object recognition. *Pattern recognition*, 32(3):453–464, 1999.

[15] Ross Girshick. Fast r-cnn. In *ICCV*, pages 1440–1448, 2015.

[16] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *CVPR*, pages 580–587, 2014.

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

[18] Khurram Azeem Hashmi, Goutham Kallempudi, Didier Stricker, and Muhammad Zeshan Afzal. Featenhancer: Enhancing hierarchical features for object detection and beyond under low-light vision. In *ICCV*, pages 6725–6735, 2023.

[19] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *ICCV*, pages 2961–2969, 2017.

[20] Mingbo Hong, Shuiwang Li, Yuchao Yang, Feiyu Zhu, Qijun Zhao, and Li Lu. Sspnet: Scale selection pyramid network for tiny person detection from uav images. *IEEE geoscience and remote sensing letters*, 19:1–5, 2021.

[21] Shih-Chia Huang, Trung-Hieu Le, and Da-Wei Jaw. Dsnet: Joint semantic learning for object detection in inclement weather conditions. *IEEE TPAMI*, 43(8):2623–2633, 2020.

[22] Hai Jiang, Ao Luo, Haoqiang Fan, Songchen Han, and Shuaicheng Liu. Low-light image enhancement with wavelet-based diffusion models. *ACM TOG*, 42(6):1–14, 2023.

[23] Hai Jiang, Ao Luo, Xiaohong Liu, Songchen Han, and Shuaicheng Liu. Lightendiffusion: Unsupervised low-light image enhancement with latent-retinex diffusion models. *arXiv preprint arXiv:2407.08939*, 2024.

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

[25] Sanket Kalwar, Dhruv Patel, Aakash Aanegola, Krishna Reddy Konda, Sourav Garg, and K Madhava Krishna. Gdip: Gated differentiable image processing for object detection in adverse conditions. pages 7083–7089. IEEE, 2023.

[26] Shuiwang Li, Yangxiang Yang, Dan Zeng, and Xucheng Wang. Adaptive and background-aware vision transformer for real-time uav tracking. In *ICCV*, pages 13989–14000, October 2023.

[27] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *ICCV*, pages 2980–2988, 2017.

[28] 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 *ECCV*, pages 740–755. Springer, 2014.

[29] Risheng Liu, Long Ma, Jiaao Zhang, Xin Fan, and Zhongxuan Luo. Retinex-inspired unrolling with cooperative prior architecture search for low-light image enhancement. In *CVPR*, pages 10561–10570, 2021.

[30] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. Ssd: Single shot multibox detector. In *ECCV*, pages 21–37. Springer, 2016.

[31] Wenyu Liu, Gaofeng Ren, Runsheng Yu, Shi Guo, Jianke Zhu, and Lei Zhang. Image-adaptive yolo for object detection in adverse weather conditions. In *AAAI*, volume 36, pages 1792–1800, 2022.

[32] Yuen Peng Loh and Chee Seng Chan. Getting to know low-light images with the exclusively dark dataset. 178:30–42, 2019.

[33] Yan-Feng Lu, Jing-Wen Gao, Qian Yu, Yi Li, Yi-Sheng Lv, and Hong Qiao. A cross-scale and illumination invariance-based model for robust object detection in traffic surveillance scenarios. *IEEE Transactions on Intelligent Transportation Systems*, 2023.

[34] Feifan Lv, Feng Lu, Jianhua Wu, and Chongsoon Lim. Mbllen: Low-light image/video enhancement using cnns. In *BMVC*, volume 220, page 4, 2018.

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

[36] Qingpao Qin, Kan Chang, Mengyuan Huang, and Guiqing Li. Denet: Detection-driven enhancement network for object detection under adverse weather conditions. In *ACCV*, pages 2813–2829, 2022.

[37] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *CVPR*, pages 779–788, 2016.

[38] Joseph Redmon and Ali Farhadi. Yolo9000: better, faster, stronger. In *CVPR*, pages 7263–7271, 2017.

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

[40] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *NeurIPS*, 28, 2015.

[41] Sebastian Ruder. An overview of gradient descent optimization algorithms. *arXiv preprint arXiv:1609.04747*, 2016.

[42] Steven A Shafer. Using color to separate reflection components. *Color Research & Application*, 10(4):210–218, 1985.

[43] Shangquan Sun, Wenqi Ren, Tao Wang, and Xiaochun Cao. Rethinking image restoration for object detection. *NeurIPS*, 35:4461–4474, 2022.

[44] Zhi Tian, Chunhua Shen, Hao Chen, and Tong He. Fcos: Fully convolutional one-stage object detection. In *ICCV*, pages 9627–9636, 2019.

[45] Biao Wang, Weifeng Li, Wenming Yang, and Qingmin Liao. Illumination normalization based on weber's law with application to face recognition. *IEEE Signal Processing Letters*, 18(8):462–465, 2011.

[46] Xiaogang Xu, Ruixing Wang, and Jiangbo Lu. Low-light image enhancement via structure modeling and guidance. In *CVPR*, pages 9893–9903, 2023.

[47] Shuzhou Yang, Moxuan Ding, Yanmin Wu, Zihan Li, and Jian Zhang. Implicit neural representation for cooperative low-light image enhancement. In *ICCV*, pages 12918–12927, 2023.

[48] 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.

[49] 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.

[50] Hongkai Zhang, Hong Chang, Bingpeng Ma, Naiyan Wang, and Xilin Chen. Dynamic r-cnn: Towards high quality object detection via dynamic training. In *ECCV*, pages 260–275. Springer, 2020.

[51] Kaipeng Zhang, Zhanpeng Zhang, Zhifeng Li, and Yu Qiao. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE signal processing letters*, 23(10):1499–1503, 2016.

[52] Shifeng Zhang, Cheng Chi, Yongqiang Yao, Zhen Lei, and Stan Z Li. Bridging the gap between anchor-based and anchor-free detection via adaptive training sample selection. In *CVPR*, pages 9759–9768, 2020.

[53] Yonghua Zhang, Jiawan Zhang, and Xiaojie Guo. Kindling the darkness: A practical low-light image enhancer. In *ACM MM*, pages 1632–1640, 2019.

A Appendix / supplemental material

Derivation of IIM. Referring to the Eq. 5 in the main text, the IIM defines a feature extracted from neighboring pixels. Consider a convolutional kernel $\mathcal{W}^{k\times k}$, where k represents the kernel size. Here, p_i and w_i denote a pixel position and its associated weight within the kernel $\mathcal{W}$, with i ranging from 1 to k^2.

$$M_{rb} = \prod_{i=1}^{k^2-1} \left(\frac{R_{p_i}}{B_{p_i}}\right)^{w_i} \left(\frac{B_{p_{i+1}}}{R_{p_{i+1}}}\right)^{w_{i+1}} = \prod_{i=1}^{k^2} \left(\frac{R_{p_i}}{B_{p_i}}\right)^{w_i} \tag{10}$$

$$\begin{aligned} log(M_{rb}) &= \sum_{i=1}^{k} w_i log(R_{p_i}) - \sum_{i=1}^{k} w_i log(B_{p_i}) \\ &= \sum_{i=1}^{k} w_i (log(m(\vec{n_{p_i}}, \vec{l_{p_i}})) + log(e^{R_{p_i}}(\lambda)) + log(\rho^{R_{p_i}}(\lambda))) \\ &- \sum_{i=1}^{k} w_i (log(m(\vec{n_{p_i}}, \vec{l_{p_i}})) + log(e^{B_{p_i}}(\lambda)) + log(\rho^{B_{p_i}}(\lambda))) \\ &= \sum_{i=1}^{k} w_i log(e^{R_{p_i}}(\lambda)) - \sum_{i=1}^{k} w_i log(e^{B_{p_i}}(\lambda)) \\ &+ \sum_{i=1}^{k} w_i log(\rho^{R_{p_i}}(\lambda)) - \sum_{i=1}^{k} w_i log(\rho^{B_{p_i}}(\lambda)) \end{aligned} \tag{11}$$

To eliminate the e term, it is imperative to adhere to the following constraints::

$$\begin{cases} \sum_{i}^{k^2} w_i log(e^{R_{p_i}}(\lambda)) = 0 \\ \sum_{i}^{k^2} w_i log(e^{B_{p_i}}(\lambda)) = 0 \end{cases} \tag{12}$$

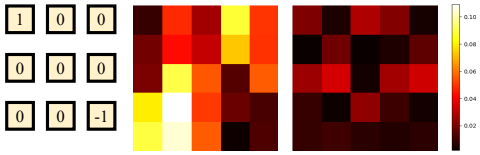
Assuming that all pixels in a given convolutional kernel are neighboring pixels, we obtain $e^{R_{p_i}} \approx e^{R_{p_j}}$, where $i, j = 1, 2, \cdots, k^2, j \neq i$. The above constraints can be equivalently expressed as $\sum_{i}^{k^2} w_i = 0$

Illumination Invariant Loss. As mentioned in Sec. 4.6.3, to optimally constrain the kernel learning process and harness the full potential of illumination-invariant information, we further employ a consistency loss, denoted as Illumination Invariant Loss (II Loss). This loss function is specifically designed to align features extracted from pairs of images taken under different lighting conditions. The fundamental concept of the II Loss is to guarantee consistency in the features extracted from these images, regardless of the variations in illumination. This is achieved by leveraging a luminance transformation function σ to adjust the illuminations, as defined as follows:

$$L = \begin{cases} \frac{1}{2}(f_{\mathcal{W}_i}(I) - f_{\mathcal{W}_i}(\sigma(I)))^2 & |f_{\mathcal{W}_i}(I) - f_{\mathcal{W}_i}(\sigma(I))| \leq \beta \\ |f_{\mathcal{W}_i}(I) - f_{\mathcal{W}_i}(\sigma(I))| - \frac{1}{2}\beta, & \text{otherwise.} \end{cases} \tag{13}$$

In our experiments, we use the gamma transformations as the function for the function σ, setting β empirically to 1, and scaling the II Loss to 0.01 of the other losses.

As discussed in Sec. 3.1, we obtain illumination-invariant features by assuming neighboring pixels exhibit high similarity of illumination. Specifically, illumination items can be factored out by performing the subtraction among the neighboring pixels, which is accomplished by imposing the zero mean constraint on the convolutional kernels. However, to eliminate the illumination term ideally, it is necessary for the average value of adjacent positions within the kernel to approach zero. The sole constraint of a zero mean does not guarantee that the illumination elimination occurs strictly between adjacent pixels; it can occur between distant pixels as well. For instance, Fig. 5(a) presents an example of a convolutional kernel that satisfies the zero-mean constraint. Even though this kernel has a zero mean, it fails to extract illumination-invariant features due to the relatively large spatial



(a) Trivial Weight (b) w/o II loss (c) w/ II loss

Figure 5: Illustration of a trivial case (a), and visualization of performing 3×3 mean filtering on the kernel weights guided by with (b) and without (c) II loss.

1)	Dataset	IIM	II-Loss	$\mathcal{K}_s$	YOLOv3	TOOD
2)	Exdark			3	71.0	72.5
3)		✓		3	71.1	74.8
4)		✓	✓	3	**72.7**	75.0
5)		✓		5	71.5	75.0
6)		✓	✓	5	**72.7**	**75.2**
7)	DarkFace			3	60.0	62.1
8)		✓		3	61.0	66.9
9)		✓	✓	3	**61.5**	**67.4**
10)		✓		5	60.2	65.8
11)		✓	✓	5	60.7	67.1

Figure 6: Ablation study of YOLOv3-based and TOOD-based detectors on ExDark and UG^2+DARK FACE datasets, where $\mathcal{K}_s$ denotes the kernel size within IIM.

Method	Bicycle	Boat	Bottle	Bus	Car	Cat	Chair	Cup	Dog	Motorbike	People	Table	mAP_{50}
Baseline	79.8	72.1	70.9	82.8	79.5	64.4	67.6	70.6	79.5	62.4	77.7	44.2	71.0
MBLLEN [34]	77.5	72.5	70.2	80.7	80.6	65.0	65.2	70.6	77.9	64.9	77.3	41.8	70.3
KIND [53]	80.2	74.4	71.5	81.0	80.3	62.2	61.3	67.5	75.8	62.1	75.9	40.9	69.4
Zero-DCE [17]	81.8	74.6	70.1	86.3	79.5	61.0	66.2	71.7	78.4	62.9	77.3	43.1	71.1
EnlightenGAN [24]	81.1	74.2	69.8	83.3	78.3	63.3	65.5	69.3	75.3	62.5	76.7	41.0	70.0
RUAS [29]	76.4	69.2	62.7	77.3	74.9	59.0	64.3	64.8	73.1	55.8	71.5	38.8	65.7
SCI [35]	80.3	74.2	73.6	82.8	78.4	64.4	65.8	71.3	78.1	62.7	78.2	42.4	71.0
NeRCo [47]	80.8	73.6	66.3	81.3	75.6	62.8	62.5	67.7	75.6	61.8	75.1	39.0	68.5
SMG [46]	78.1	72.1	65.8	81.6	78.3	63.7	64.5	67.6	76.3	57.4	73.7	42.4	68.5
DENet [36]	81.1	75.0	73.9	87.1	79.7	63.5	66.3	69.6	76.3	61.4	76.7	44.9	71.3
PENet [49]	76.5	71.9	67.4	84.2	78.0	59.9	64.6	66.7	74.8	62.5	73.9	45.1	68.8
MAET [7]	81.5	73.7	74.0	88.2	80.9	68.8	66.9	71.8	79.3	60.2	78.8	46.3	72.5
Ours	82.4	74.0	72.7	85.4	81.0	67.2	66.5	71.5	81.8	65.2	78.6	45.7	72.7

Table 6: Quantitative comparisons of the ExDark dataset based on YOLOv3 detector.

separation between the positive and negative positions. Unfortunately, as the convolutional kernel size increases, this issue becomes more pronounced and leads to a degradation in performance. To this end, the II Loss is proposed to encourage consistency of outputs from IIM across images with different illuminations, preventing trivial solutions within the kernel implicitly. As shown in Fig. 5(b)(c), we visualize the 5×5 kernel with and without the II Loss. For each position in the 5×5 kernel, we compute the mean value using a sliding window of size 3×3. It can be observed that without the II Loss, the local means within the kernel do not tend towards zero, indicating that the features extracted using this kernel are may not illumination-invariant. In contrast, when the II Loss is applied, the local values of the kernel are significantly constrained. To further validate the effectiveness of the II Loss, we present the results of the II Loss on different convolutional kernel sizes within IIM in rows 3$\sim$6 and 8$\sim$11 of Table 6. By comparing the performance gains of different convolutional kernel sizes, we can see that the larger kernel sizes lead to more significant improvements. This strongly suggests the effectiveness of our II Loss in constraining the degrees of freedom in the kernel.

Detailed Results on ExDark. In this section, we report the average precision for each category of the ExDark dataset as shown in Table 6 and 7. Note that, we further introduce more advanced LLIE methods, including Zero-DCE [17], EnlightenGAN [24], SCI [35], and NeRCo [47] based on YOLOv3 and TOOD. Unfortunately, despite the outstanding performance exhibited by these LLIE methods in image restoration tasks, they struggle with effectively enhancing specific downstream tasks. For example, the state-of-the-art LLIE method, NeRCo, exhibits the worst performance compared to other LLIE methods. This phenomenon further proves the existence of the gap between optimization goals for image restoration and object detection tasks. Additionally, compared to end-to-end approaches, such cascade paradigms limit the potential for deploying these LLIE-based low-light detection techniques to practical applications.

YOLA vs. FeatEnHancer. For a fair comparison, we follow the FeatEnHancer's [18] experimental setting to implement the RetinaNet [27]-based detectors as shown in Table 8. We can see that even though our baseline implementation on the ExDark dataset is inferior to FeatEnHancer's, the

Method	Bicycle	Boat	Bottle	Bus	Car	Cat	Chair	Cup	Dog	Motorbike	People	Table	mAP_{50}
Baseline	80.6	75.8	71.1	88.1	76.8	70.4	66.8	69.2	85.4	61.5	76.1	48.2	72.5
MBLLEN [34]	80.8	77.8	72.8	89.3	78.7	73.5	67.5	69.4	85.2	62.9	77.3	47.2	73.5
KIND [53]	81.7	77.7	70.3	88.4	78.1	69.7	67.2	67.8	84.1	61.6	76.6	47.8	72.6
Zero-DCE [17]	81.8	79.0	72.9	89.6	77.9	71.9	68.5	69.8	84.8	62.9	78.0	49.5	73.9
EnlightenGAN [24]	80.7	77.6	70.4	88.8	76.9	70.6	67.9	68.7	84.4	62.2	77.5	49.6	73.0
RUAS [29]	78.4	74.3	67.4	85.1	72.4	67.7	67.3	65.2	77.9	56.1	73.4	47.0	69.4
SCI [35]	81.3	78.1	71.6	89.4	77.6	71.1	68.0	70.9	85.0	63.0	77.2	49.2	73.5
NeRCo [47]	78.8	75.6	70.8	87.6	75.7	69.1	66.8	69.5	82.5	59.9	76.0	49.3	71.8
SMG [46]	78.2	75.9	69.9	87.3	75.1	71.3	66.5	67.2	84.2	60.1	75.1	46.7	71.5
DENet [36]	80.9	78.2	70.9	88.3	77.5	71.6	67.2	70.3	87.3	62.0	77.3	49.9	73.5
PENet [49]	76.0	72.3	66.7	84.4	72.2	65.4	63.3	65.8	79.1	53.1	71.0	44.6	67.8
MAET [7]	80.5	77.3	74.0	90.1	78.3	73.4	69.6	70.7	86.6	64.4	77.6	48.5	74.3
Ours	83.9	78.7	75.3	88.8	79.0	73.4	69.9	71.9	86.8	66.3	78.3	49.8	75.2

Table 7: Quantitative comparisons of the ExDark dataset based on TOOD detector.

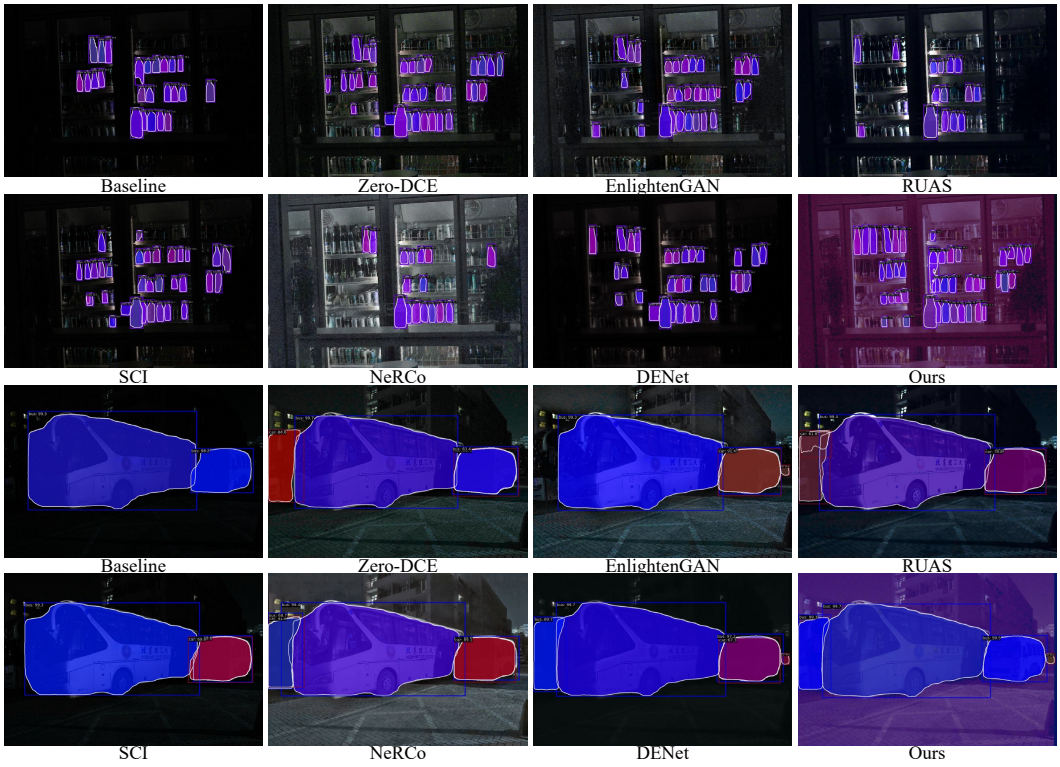


Figure 7: Qualitative comparisons of Mask R-CNN-based detector on LIS dataset. Our YOLA shows more comprehensive segmentation and detection results, with an increased number of bottles detected (top 2 rows) and successful recognition of the challenging car (right side of the bottom 2 rows). Best viewed with zooming in.

integration of YOLA enables our method to achieve the best performance (1.9 mAP significant improvement compared to baseline). For UG^2+DARK FACE dataset, FeatEnHancer decreases the baseline performance by 0.1 mAP, which is attributed to hierarchical features that failed to be captured by RetinaNet, as claimed in [18]. In contrast, our YOLA, triggered from the physics-based model perspective without elaborate design, surpassing the baseline with a remarkable improvement of 2.5 mAP. It strongly suggests the generalizability and effectiveness of YOLA.

More Visualization. In Fig. 7 and 8, additional visual results are presented, showcasing selected challenging cases. In comparison to other methods, our method exhibits superior recall and more precise segmentation performance under extreme low-light conditions.

YOLA on Lowl-light Instance Segmentation. To further explore YOLA's capabilities, we also evaluate YOLA in the low-light instance segmentation tasks. We report the quantitative comparisons

Dataset	Method	mAP_{50}
Exdark	Baseline	72.1
	w/ FeatEnHancer [18]	72.6(+**0.5**)
	Baseline†	70.9
	w/ YOLA	**72.8**(+**1.9**)
DarkFace	Baseline	47.3
	w/ FeatEnHancer [18]	47.2(-**0.1**)
	Baseline†	50.2
	w/ YOLA	**52.7**(+**2.5**)

Table 8: Quantitative comparisons (YOLA vs. FeatEnHancer) of ExDark and UG^2+DARK FACE datasets based on RetinaNet. Red and blue colors represent **improvement** and **degradation** of performance, respectively, compared to the baseline. † indicates our implemented baseline.

Method	AP^{seg}	AP^{seg}_{50}	AP^{seg}_{75}	AP^{box}	AP^{box}_{50}	AP^{box}_{75}
Baseline	34.2	55.6	34.7	41.3	63.9	44.6
DENet [36]	38.6	61.7	39.8	46.4	70.1	51.0
PENet [49]	36.1	58.8	36.4	43.6	67.3	47.1
Zero-DCE [17]	38.7	62.0	39.0	46.4	70.0	50.9
EnlightenGAN [24]	38.4	61.5	39.2	45.8	69.5	49.7
RUAS [29]	36.1	58.6	36.4	43.8	66.7	48.0
SCI [35]	36.5	59.5	37.0	44.3	67.3	48.4
NeRCo [47]	36.7	60.3	38.6	44.6	68.3	48.6
SMG [46]	37.4	60.3	38.7	44.7	67.4	49.2
Ours	**39.8**	**63.5**	**41.4**	**47.5**	**70.9**	**51.8**

Table 9: Quantitative comparisons of the LIS dataset based on Mask RCNN, where AP^{seg} and AP^{box} indicate the average precision of segmentation and detection, respectively.

Figure 8: Qualitative comparisons of Mask R-CNN-based detector on LIS dataset. Our YOLA outperforms LLIE-based and low-light object detection methods. Best viewed with zooming in.

of several advanced LLIE and low-light object methods using Mask R-CNN [19] on the low-light instance segmentation (LIS) [5] dataset, as shown in Table 9. We can see that our YOLA achieves the best performance across all metrics, indicating that YOLA not only facilitates low-light object detection but also low-light instance segmentation.

NeurIPS Paper Checklist

The checklist is designed to encourage best practices for responsible machine learning research, addressing issues of reproducibility, transparency, research ethics, and societal impact. Do not remove the checklist: **The papers not including the checklist will be desk rejected.** The checklist should follow the references and follow the (optional) supplemental material. The checklist does NOT count towards the page limit.

Please read the checklist guidelines carefully for information on how to answer these questions. For each question in the checklist:

- You should answer [Yes] , [No] , or [NA] .
- [NA] means either that the question is Not Applicable for that particular paper or the relevant information is Not Available.
- Please provide a short (1–2 sentence) justification right after your answer (even for NA).

The checklist answers are an integral part of your paper submission. They are visible to the reviewers, area chairs, senior area chairs, and ethics reviewers. You will be asked to also include it (after eventual revisions) with the final version of your paper, and its final version will be published with the paper.

The reviewers of your paper will be asked to use the checklist as one of the factors in their evaluation. While "[Yes] " is generally preferable to "[No] ", it is perfectly acceptable to answer "[No] " provided a proper justification is given (e.g., "error bars are not reported because it would be too computationally expensive" or "we were unable to find the license for the dataset we used"). In general, answering "[No] " or "[NA] " is not grounds for rejection. While the questions are phrased in a binary way, we acknowledge that the true answer is often more nuanced, so please just use your best judgment and write a justification to elaborate. All supporting evidence can appear either in the main paper or the supplemental material, provided in appendix. If you answer [Yes] to a question, in the justification please point to the section(s) where related material for the question can be found.

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: See abstract and introduction.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [NA]

   Justification: **[TODO]**

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.

- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: We include the assumptions in Section 3.1, and complete proofs of IIM in appendix.

   Guidelines:

   - The answer NA means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: See section 4.1

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example

(a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
(b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
(c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: This work only use public datasets, and we provide the code in supplemental materials.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: See section 4.1.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [NA]

Justification: **[TODO]**

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: See the experiment section.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: **[TODO]**

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [NA]

    Justification: **[TODO]**

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: **[TODO]**

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: See section 4.1.

    Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification: **[TODO]**

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [No]

    Justification: **[TODO]**

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [No]

    Justification: **[TODO]**

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.