FeatEnHancer: Enhancing Hierarchical Features for Object Detection and Beyond Under Low-Light Vision

Khurram Azeem Hashmi$^{1,2}$, Goutham Kallempudi$^2$, Didier Stricker$^{1,2}$ and Muhammamd Zeshan Afzal$^{1,2}$

$^1$DFKI - German Research Center for Artificial Intelligence, $^2$RPTU Kaiserslautern, {khurram.azeem.hashmi, didier.stricker, muhammad.zeshan.afzal}@dfki.de, kallempu@rptu.de

Figure 1: Learned hierarchical representation and enhanced image from ourFeatEnHancer. We train our FeatEnHancer on a downstream object detection task and visualize these images from the validation set. These maps and enhanced images show that despite producing less visually appealing images, our model enhances task-related features. Best viewed on the screen.

Abstract

Extracting useful visual cues for the downstream tasks is especially challenging under low-light vision. Prior works create enhanced representations by either correlating visual quality with machine perception or designing illumination-degrading transformation methods that require pre-training on synthetic datasets. We argue that optimizing enhanced image representation pertaining to the loss of the downstream task can result in more expressive representations. Therefore, in this work, we propose a novel module, FeatEnHancer, that hierarchically combines multiscale features using multi-headed attention guided by task-related loss function to create suitable representations. Furthermore, our intra-scale enhancement improves the quality of features extracted at each scale or level, as well as combines features from different scales in a way that reflects their relative importance for the task at hand. FeatEnHancer is a general-purpose plug-and-play module and can be incorporated into any low-light vision pipeline. We show with extensive experimentation that the enhanced representation produced with FeatEnHancer significantly and consistently improves results in several low-light vision tasks, including dark object detection (+5.7 mAP on ExDark), face detection (+1.5 mAP on DARK FACE), nighttime semantic segmentation (+5.1 mIoU on ACDC), and video object detection (+1.8 mAP on DarkVision), highlighting the effectiveness of enhancing hierarchical features under low-light vision.

1. Introduction

Recent remarkable advancements in high-level vision tasks have shown that given a high-quality image, current vision backbone networks [20, 15, 12, 32, 31], object detectors [42, 28, 43, 19, 2, 3, 49, 4, 71, 64, 65] and semantic segmentation models [34, 48, 57, 7, 58] can effectively learn desired features to perform vision tasks. Similarly, modern low-light image enhancement (LLIE) methods [44, 67, 14, 21, 17, 25] are capable of transforming a low-light image into a visual-friendly representation. However, a naive combination of the two brings sub-optimal gains when it comes to high-level vision tasks under low-light vision.
This work explores the underlying reasons for the low performance of the combination of LLIE with high-level vision methods and observes the following limitations: 1) Although existing LLIE methods push the envelope of visual perception for human eyes, they do not align with vision backbone networks [20, 12, 15, 32, 31] due to lack of multi-scale features. For instance, it is likely that the enhancement method increases brightness in some regions. However, it simultaneously corrupts the edges and texture information of objects. 2) The pixel distribution among different low-light images may have huge variance owing to the disparity in less illuminated environments [17, 25, 68]. This increases intra-class variance in some cases (see Fig. 3, where only one bicycle is recognized by [17] instead of two bicycles in the ground-truth). 3) Current LLIE approaches [14, 17, 25, 21, 44, 56, 67] employ enhancement loss functions to optimize the enhancement networks. These loss functions compel the network to attend to all pixels equally, lacking the learning of informative details necessary for high-level downstream vision tasks such as object pose and shape for object detection. Furthermore, to train these enhancement networks, most of them [44, 14, 67, 56] require a set of high-quality images, which is hardly available in a real-world setting.

Motivated by these observations and inspired by recent developments in LLIE [17, 25] and vision-based backbone networks [15, 32, 31], this paper aims to bridge the gap by exploring an end-to-end trainable recipe that jointly optimizes the enhancement and downstream task objectives in a single network. To this end, we present FeatEnHancer, a general-purpose feature enhancer that learns to enrich multi-scale hierarchical features favourable for downstream vision tasks in a low-light setting. An example of learned hierarchical representation and the enhanced image is illustrated in Fig. 1.

In particular, our FeatEnHancer first downsamples a low-light RGB input image to construct multi-scale hierarchical representations. Subsequently, these representations are fed to our Feature Enhancement Network (FEN), which is a deep convolutional network, employed to enrich intra-scale semantic representations. Note that the parameters of FEN can be adjusted through task-related loss functions, which pushes the FEN to only enhance the task-related features. This multi-scale learning allows the network to enhance both global and local information from higher and lower-resolution features, respectively. Once the enhanced representations on different scales are obtained, the remaining obstacle is to fuse them effectively. To achieve this, we select two different strategies to capture both global and local information from higher and lower-resolution features. First, to merge high-resolution features, inspired by multi-head attention in [50], we design a Scale-aware Attentional Feature Aggregation (SAFA) method that jointly attends information from different scales. Second, for lower-resolution features, the skip connection [20] scheme is adopted to merge the enhanced representation from SAFA to lower-resolution features. With these jointly learned hierarchical features, our FeatEnHancer provides semantically powerful representations which can be exploited by advanced methods such as feature pyramid networks [27] for object detection [43] and instance segmentation [19], or UNet [45] for semantic segmentation [34].

The main contributions of this work can be summarized as follows:

1. We propose FeatEnHancer, a novel module that enhances hierarchical features to boost downstream vision tasks under low-light vision. Our intra-scale feature enhancement and scale-aware attentional feature aggregation schemes are aligned with vision backbone networks and produce powerful semantic representations. FeatEnHancer is a general-purpose plug-and-play module that can be trained end-to-end with any high-level vision task.

2. To the best of our knowledge, this is the first work that fully exploits multi-scale hierarchical features in low-light scenarios and generalizes to several downstream vision tasks such as object detection, semantic segmentation, and video object detection.

3. Extensive experiments on four different downstream vision tasks covering both images and videos demonstrate that our method brings consistent and significant improvements over baselines, LLIE methods and task-specific state-of-the-art approaches.

2. Related Work

2.1. Enhancing Low-Light Images

Deep learning-based LLIE methods focus on improving the visual quality of low-light images that satisfies human visual perception [23, 22]. Most LLIE approaches [14, 44, 56, 67] operate under a supervised learning paradigm, requiring paired data during training. Unsupervised GAN-based methods [21] eliminate the need for paired data during the training. However, their performance relies on the careful choice of unpaired data. Recently, zero-reference methods [17, 25, 68] discard the need for both paired and unpaired data to enhance low-light images by designing a set of non-reference loss functions. Inspired by these recent developments, this work aims to bridge low-light enhancement and downstream vision tasks (such as object detection [10, 33, 62], semantic segmentation [60, 47], and video object detection [63]) by enhancing multi-scale hierarchical features without needing paired or unpaired data to boost performance.
2.2. Enhancing Low-Light for Downstream Vision Tasks

These approaches consider machine perception as the criteria for success while enhancing images to improve downstream vision tasks. One obvious way to achieve this goal is to apply the LLIE methods as an initial step [70, 17]. However, this leads to unsatisfactory results (see Table 2, 4, and 5). Recently, another line of work has explored end-to-end pipelines, optimizing both enhancement and individual tasks during training, and our work follows the same spirit.

Face detection. Liang et al. [26] propose an effective information extraction scheme from low-light images by exploiting multi-exposure generation. Furthermore, bi-directional domain adaptation [52, 51] and parallel architecture that jointly performs enhancement and detection [37] are presented to advance the research. However, these approaches are carefully designed to tackle face detection [62, 53] only and deliver minor improvements when applied to generic object detection [51]. Contrarily, our FeatEnHancer is a general-purpose module. It significantly improves several downstream vision tasks. Hence, we refrain from comparing our method to architectures only evaluated for face detection.

Dark object detection. Dark (low-light) object detection [10, 30] methods have emerged recently, thanks to the real-world low illumination datasets [33, 39]. IA-YOLO [30] introduces a convolutional neural network (CNN)-based parameter predictor that learns the optimal configuration for the filters employed in the differential image processing module. Most related to our work is MAET [10], which investigates the physical noise model and image signal processing (ISP) pipeline under low illumination and learns the model to predict degradation parameters and object features. To avoid feature entanglement, they impose orthogonal tangent regularity to penalize cosine similarity between objects and degrading features. However, owing to the weather-specific hyperparameters in [30] and degradation parameters in [10], these works rely on large synthetic datasets to achieve desired performance. Unlike them, our FeatEnHancer is optimized from the task-related loss functions and does not require any pre-training on synthetic datasets mimicking low-light or harsh weather conditions.

Other high-level vision tasks. Besides face and object detection, recent research has explored high-level computer vision tasks like semantic segmentation [6, 34]. Xue et al. [60] devise a contrastive-learning strategy to improve visual and machine perception simultaneously, achieving impressive performance on nighttime semantic segmentation of adverse conditions dataset with correspondences (ACDC) dataset [47]. Furthermore, DarkVision [63] has emerged recently to tackle video object detection under low-light vision. In this work, thanks to [47, 63], we apply FeatEnHancer to semantic segmentation and video object detection under low-light vision to investigate its generalization capabilities.

2.3. Learning Multi-scale Hierarchical Features

Representing objects at varying scales is one of the main difficulties in computer vision. Therefore, the work in this domain goes back to the era of hand-engineered features [36, 11, 38, 24]. Modern object detectors [43, 28, 2, 71, 49, 40, 65] exploit multi-scale features to tackle this challenge. Similarly, multi-scale representations [34] and pyramid pooling schemes [69] have been proposed for effective semantic segmentation. Moreover, current improvements in vision-based backbone networks [15, 31, 32] demonstrate that learning hierarchical features during feature extraction directly uplifts the downstream vision tasks [19, 57, 2]. However, the multi-scale and hierarchical structures of CNN have not been fully explored for low-light vision tasks.

Under harsh weather conditions, DENet [41] employs Laplacian Pyramid [1] to decompose images into low and high-frequency components for object detection. Despite the encouraging results, the multi-scale feature learning in DENet relies on the Laplacian pyramid, which is susceptible to noise and may produce inconsistencies in regions with high contrast or sharp edges. Alternatively, aligned with the multi-scale learning in modern vision backbone networks [27, 32, 31], our FeatEnHancer employs CNN to generate multi-scale feature representations, which are fused through the scale-aware attentional feature aggregation and skip connections. Our approach is much more flexible and aligns with downstream vision tasks, boosting state-of-the-art results on multiple downstream vision tasks.

3. Proposed Approach

The key idea of this paper is to design a general-purpose pluggable module that strengthens machine perception under low-light vision to solve several downstream vision tasks such as object detection, semantic segmentation, and video object detection. The overall architecture of FeatEnHancer is exhibited in Fig. 2. Our FeatEnHancer takes a low-light image as input and adaptively boosts its semantic representation by enriching task-related hierarchical features. We now discuss the key components of FeatEnHancer in detail.

3.1. Hierarchical Feature Enhancement

Inspired by the recent improvements in vision-based backbone networks [15, 31, 32], we introduce the enhancement of hierarchical features through jointly optimizing feature enhancement and downstream tasks under low-light vision. Unlike [15, 31, 32], our goal is to extract spatial features from low-light images and generate meaningful semantic representations. In order to enhance hierarchical features, we first construct multi-scale representations from the low-light input image. Later, we feed these multi-scale
representations to our feature enhancement network.

**Constructing multi-scale representations.** We take a low-light RGB image $I \in \mathbb{R}^{H \times W \times C}$ as input and employ regular convolutional operator $\text{Conv}()$ on $I$ to generate $I_q \in \mathbb{R}^{\frac{H}{4} \times \frac{W}{4} \times 3}$ and $I_o \in \mathbb{R}^{\frac{H}{8} \times \frac{W}{8} \times 3}$ representing the quarter and octa scale of an input image, respectively. To summarize, it can be written as:

$$
\begin{align*}
I_q &= \text{Conv}(I) \\
I_o &= \text{Conv}(I_q)
\end{align*}
$$

where $K$ and $S$ denote kernel size and stride, and $H$, $W$, and $C$ represent the height, width, and channels of an image.

**Feature enhancement network.** In order to enhance features at each scale, we require an enhancement network that learns to enhance spatial information important for downstream tasks. Inspired by low-light image enhancement networks [17, 25], we design a fully convolutional intra-scale feature extraction network (FEN). However, unlike [17, 25], our FEN introduces a single convolutional layer at the beginning that generates a feature map $F \in \mathbb{R}^{H \times W \times C}$, where $C$ is transformed from 3 to 32 by keeping the resolution $(H \times W)$ same as the input. Then a series of six convolutional layers with symmetrical skip concatenation is applied, where each convolutional layer, with $K = 3$ and $S = 1$, is accompanied by the ReLU activation function. We apply FEN on each scale $I$, $I_q$, and $I_o$ separately, and obtain multi-scale feature representations, denoted as $F_q$, $F_o$, and $F_o$, respectively. This multi-scale learning allows the network to enhance both global and local information from higher and lower-resolution features. Hence, we ignore down-sampling and batch normalization to preserve semantic relations between neighbouring pixels which is similar to [17]. However, we discard the last convolutional layer of DCENet [17] in our FEN and propagate the final enhanced feature representations from each scale for the multi-scale feature fusion. Note that the implementation details of FEN in FeatEnHancer are independent of the proposed module, and even more, advanced image enhancement networks such as [68] can be applied to improve performance. Now, we discuss multi-scale feature fusion in detail.

### 3.2. Multi-scale Feature Fusion

Since we already have multi-scale feature representations $(F_q, F_o, F_o)$ from FEN, the remaining obstacle is to fuse them effectively. Lower-scale features $(F_q)$ capture more abstract information, such as shapes and patterns. Therefore, naive aggregation leads to inferior performance (see Table 6a). Hence, we adopt two different strategies to capture both global and local information from higher and lower-resolution features. First, inspired by multi-head attention in [50] that enables the network to jointly learn information from different channels, we design a scale-aware attentional feature aggregation (SAFA) module that jointly attends to features from different scales. Second, we adopt a skip connection [20] (SC) scheme to integrate low-level information from $F_o$ and the enhanced representation from SAFA to obtain the final enhanced hierarchical representation. Adopting SAFA for merging high-resolution features and SC for lower-resolution features leads to a more robust hierarchical representation (see Table 6b). Now, we discuss SAFA in detail.

**Scale-aware Attentional Feature Aggregation.** Even though high-resolution features assist in capturing fine de-
where $W$ and $U$ are the attentional weights. Later, $Q$ and $K$ are concatenated to form the set of hierarchical features $F_{q+k}$, which are split into $N$ blocks along the channel dimension $C$:

$$\bar{F}_{q+k} = F_{q+k}[; ; , (n-1) \frac{C}{N} : n \frac{C}{N}],$$

where $n \in \{1, 2, ..., N\}$ and $N$ is the total number of attentional blocks. The $F_{q+k} \in \mathbb{R}^{H \times W \times C}$ is used to compute attention weights $W$ in a single attention block as follows:

$$W_{q+k} = F_{q} \cdot F_{k}^{\top},$$

$$\bar{W}_{q+k} = \frac{\exp(W_{q+k})}{\sum_{l=1}^{L} \exp(W_{q+k})},$$

where $W_{q+k}$ is the attention weights of $F_{q}$ and $F_{k}$ for $n$-th block, and $\bar{W}_{q+k}$ is the normalized form of $W_{q+k}$. Derived from the $n$-th block of normalized attention weights, we apply weighted sum to compute the $n$-th block of enhanced hierarchical representation $\bar{F}_{h} \in \mathbb{R}^{H \times W \times C}$ as follows:

$$\bar{F}_{h} = \sum_{l=1}^{L} \bar{W}_{q+k} \cdot F_{q+k},$$

now we concatenate all $\bar{F}_{h}$ along the channel dimension to obtain $\bar{F}_{h} \in \mathbb{R}^{H \times W \times C}$. Note that although $\bar{F}_{h}$ is the same size as $Q$ and $K$, it contains far richer representations, encompassing information from multi-scale high-resolution features.

Subsequently, as explained earlier in Sec. 3.2, with the help of skip connections (SC), we integrate $F_h$ and $\bar{F}_h$ to obtain the final enhanced hierarchical representation covering both global and local features, as illustrated in Figure 1 and 2. Note that prior to the skip connection, we upsample $\bar{F}_h$ and $F_h$, where the upsampling operation $U(\cdot) \in \mathbb{R}^{H \times W \times C} \rightarrow \mathbb{R}^{H \times W \times C}$ is performed with simple bi-linear interpolation operation, which is much faster than using transposed convolutions [13]. Unlike image enhancement in existing works [17, 25, 10], with a multi-scale hierarchical feature enhancement strategy, our FeatEnHancer learns a powerful semantic representation by capturing both local and global features. This makes it a general-purpose module to enhance hierarchical features, boosting machine perception under low-light vision.

### 4. Experiments

We conduct extensive experiments for evaluating the proposed FeatEnHancer module to several downstream tasks under the low-light vision, including generic object detection [33, 39], face detection [62], semantic segmentation [47], and video object detection [63]. Table 1 summarizes the crucial statistics of the employed datasets. This section first compares the proposed method with powerful baselines, existing LLIE approaches, and task-specific state-of-the-art methods. Then, we ablate the important design choices of our FeatEnHancer. We provide complete implementation details for each experiment in Appendix A.

#### 4.1. Dark Object Detection

**Settings.** For dark object detection experiments on the real-world data, we consider the exclusively dark (ExDark) [33] dataset (see Table 1). We adopt RetinaNet [28] as a typical detector and Featurized Query R-CNN [65] (FQ R-CNN) as an advanced object detection framework to report results. In the case of both detectors, pre-trained models on COCO [29] are fine-tuned on each dataset. For RetinaNet, images are resized to 640×640, and we train the network using 1×schedule in mmdetection [5] (12 epochs using SGD optimizer [46] with an initial learning rate of 0.0001). For Featurized Query R-CNN, we employ multi-scale training [4, 49, 65] (shorter side ranging from 400 to 800 with a longer side of 1333). The FQ R-CNN is trained for 50000 iterations using ADAMW [35] optimizer (initial learning rate of 0.0000025, weight decay of 0.0001, and batch size of 8). Note that for each object detection framework, we adopt the same settings while reproducing results of our work, baseline, LLIE approaches, and task-specific state-of-the-art methods.

We compare our FeatEnHancer to several state-of-the-art LLIE methods, including KIND [67], RAUS [44], EnGAN [21], MBILLEN [14], Zero-DCE [17], Zero-DCE++ [17], and state-of-the-art dark object detection method, MAET [10]. For LLIE methods, all images are
Methods RetinaNet FQ R-CNN

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>mAP</td>
<td>mAP</td>
<td>mAP</td>
<td>mAP</td>
<td>mAP</td>
<td>mAP</td>
<td>mAP</td>
<td>mAP</td>
<td>mAP</td>
<td>mAP</td>
<td>mAP</td>
</tr>
<tr>
<td>Baseline</td>
<td>72.1</td>
<td>64.7</td>
<td>70.7</td>
<td>70.3</td>
<td>70.4</td>
<td>70.6</td>
<td>71.0</td>
<td>71.8</td>
<td>72.6</td>
<td>72.6</td>
</tr>
<tr>
<td>fAP</td>
<td>fAP</td>
<td>fAP</td>
<td>fAP</td>
<td>fAP</td>
<td>fAP</td>
<td>fAP</td>
<td>fAP</td>
<td>fAP</td>
<td>fAP</td>
<td>fAP</td>
</tr>
<tr>
<td>mIoU</td>
<td>mIoU</td>
<td>mIoU</td>
<td>mIoU</td>
<td>mIoU</td>
<td>mIoU</td>
<td>mIoU</td>
<td>mIoU</td>
<td>mIoU</td>
<td>mIoU</td>
<td>mIoU</td>
</tr>
<tr>
<td>Baseline</td>
<td>47.3</td>
<td>42.1</td>
<td>47.2</td>
<td>47.3</td>
<td>45.1</td>
<td>47.1</td>
<td>44.3</td>
<td>47.2</td>
<td>47.3</td>
<td>47.2</td>
</tr>
<tr>
<td>RAUS [44]</td>
<td>19.9</td>
<td>17.6</td>
<td>19.8</td>
<td>20.1</td>
<td>19.3</td>
<td>19.8</td>
<td>18.7</td>
<td>19.9</td>
<td>19.9</td>
<td>19.9</td>
</tr>
<tr>
<td>KIND [67]</td>
<td>51.5</td>
<td>51.5</td>
<td>51.9</td>
<td>51.9</td>
<td>51.9</td>
<td>52.0</td>
<td>52.4</td>
<td>56.5</td>
<td>56.5</td>
<td>56.5</td>
</tr>
<tr>
<td>Zero-DCE++ [25]</td>
<td>42.6</td>
<td>42.6</td>
<td>42.6</td>
<td>42.6</td>
<td>42.6</td>
<td>42.6</td>
<td>42.6</td>
<td>42.6</td>
<td>42.6</td>
<td>42.6</td>
</tr>
<tr>
<td>EnGAN [21]</td>
<td>44.0</td>
<td>44.0</td>
<td>44.0</td>
<td>44.0</td>
<td>44.0</td>
<td>44.0</td>
<td>44.0</td>
<td>44.0</td>
<td>44.0</td>
<td>44.0</td>
</tr>
<tr>
<td>MBLLEN [14]</td>
<td>45.1</td>
<td>45.1</td>
<td>45.1</td>
<td>45.1</td>
<td>45.1</td>
<td>45.1</td>
<td>45.1</td>
<td>45.1</td>
<td>45.1</td>
<td>45.1</td>
</tr>
<tr>
<td>Zero-DCE [17]</td>
<td>45.2</td>
<td>45.2</td>
<td>45.2</td>
<td>45.2</td>
<td>45.2</td>
<td>45.2</td>
<td>45.2</td>
<td>45.2</td>
<td>45.2</td>
<td>45.2</td>
</tr>
<tr>
<td>MAET [10]</td>
<td>45.5</td>
<td>45.5</td>
<td>45.5</td>
<td>45.5</td>
<td>45.5</td>
<td>45.5</td>
<td>45.5</td>
<td>45.5</td>
<td>45.5</td>
<td>45.5</td>
</tr>
<tr>
<td>FeatEnHancer</td>
<td>46.3</td>
<td>46.3</td>
<td>46.3</td>
<td>46.3</td>
<td>46.3</td>
<td>46.3</td>
<td>46.3</td>
<td>46.3</td>
<td>46.3</td>
<td>46.3</td>
</tr>
</tbody>
</table>

Table 2: Quantitative comparison on ExDark dataset. Results obtained on the commonly used evaluation metrics are highlighted. Our FeatEnHancer brings consistent improvements and achieves new state-of-the-art results with FQ R-CNN.

Results on ExDark. Table 2 lists the results of LLIE works, MAET, and the proposed method on both object detection frameworks. It is evident that our FeatEnHancer brings consistent and significant gains over prior methods. Note that, while the performance of MAET and our method is comparable on RetinaNet (≈ 72 AP50), the proposed FeatEnHancer outperforms MAET by a significant margin on FQ R-CNN, achieving the new state-of-the-art AP50 of 86.3. Furthermore, Figure 3 shows four detection examples from our method and the two best competitors using FQ R-CNN as a detector. These results illustrate that despite the inferior visual quality, our FeatEnHancer enhances hierarchical features that are favourable for dark object detection, producing state-of-the-art results.

4.2. Face Detection on DARK FACE

Settings. The DARK FACE [53, 62] is a challenging face detection dataset released for the UG2 competition. For experiments on the DARK FACE (see Table 1), the images are resized to a larger resolution of 1500 × 1000 for all methods. We adopt the same object detection frameworks of RetinaNet and FQ R-CNN and follow identical experimental settings, as explained in Sec. 4.1.

Results. The performance of FeatEnHancer, MAET, and six LLIE methods, using RetinaNet and Featurized Query R-CNN, are summarized in Table 3. Note that a few LLIE methods [17, 25, 67] yield superior results than our approach in the case of RetinaNet. We argue that due to tiny faces with highly dark images in the DARK FACE dataset, RetinaNet fails to capture information even from the enhanced hierarchical features. We discuss this behaviour with an example in Appendix B. On the other hand, LLIE approaches directly provide well-lit images that bring slightly bigger gains (+0.1 mAP50) in this case. However, note that with the more strong detector, our FeatEnHancer surpasses all the LLIE methods and MAET by a significant margin (+1.5 mAP50), achieving mAP50 of 69.0.

4.3. Nighttime Semantic Segmentation on ACDC

Settings. We utilize nighttime images from the ACDC dataset [47] (see Table 1) to report results on semantic segmentation in a low-light setting. DeepLabV3+ [7] is adopted as the segmentation baseline from mmseg [8] for straightforward comparison with the concurrent work [60]. We follow identical experimental settings as in [60]. Refer to Appendix A for complete implementation details.

Results. We compare our method with several state-of-the-art LLIE methods, including RetinexNet [54] KIND [67], DRBN [59], FIDE [61], ZerodCE [17], SSIENet [66], Xue et al. [60], and MAET [10], and our FeatEnHancer achieves new state-of-the-art results.
Figure 4: Qualitative comparison of FeatEnHancer with previous best work [60] on the ACDC nighttime semantic segmentation. FeatEnHancer provides more accurate segmentations.

Table 5: Comparing FeatEnHancer with LLIE methods on the DarkVision dataset. FeatEnHancer is the only method that boosts the performance of the powerful baseline method on both illumination levels.

<table>
<thead>
<tr>
<th>Method</th>
<th>Illumination (3.2) mAP</th>
<th>Illumination (0.2) mAP</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline [55]</td>
<td>32.8</td>
<td>10.4</td>
</tr>
<tr>
<td>RAUS [44]</td>
<td>7.42</td>
<td>5.19</td>
</tr>
<tr>
<td>EnGAN [21]</td>
<td>7.83</td>
<td>5.41</td>
</tr>
<tr>
<td>MBLLEN [14]</td>
<td>7.82</td>
<td>5.39</td>
</tr>
<tr>
<td>KIND [67]</td>
<td>7.43</td>
<td>5.25</td>
</tr>
<tr>
<td>Zero-DCE++ [25]</td>
<td>7.51</td>
<td>5.02</td>
</tr>
<tr>
<td>Zero-DCE [17]</td>
<td>7.83</td>
<td>5.43</td>
</tr>
<tr>
<td>FeatEnHancer</td>
<td>34.6</td>
<td>11.2</td>
</tr>
</tbody>
</table>

4.4. Video Object Detection on DarkVision

**Settings.** We extend our experiments from static images to video domain to test the generalization capabilities of our method. The video object detection under low-light vision is evaluated on the recently emerged DarkVision dataset [63] (see Table 1 for dataset details). Although the dataset is not publicly available yet, we sincerely thank the authors of [63] for providing prompt access. To evaluate our FeatEnHancer under low light settings, we take the low-end camera split on two different illumination levels, i.e., 0.2 and 3.2. For ablation studies, we adopt a 3.2% illumination level split. We consider SELSA [55] as our baseline and follow identical experimental settings with the ResNet-50 backbone network in the mmtracking [9]. To establish a direct comparison, we enhance all video frames first through LLIE methods and feed these frames to the baseline, as done in Sec. 4.1. As a common practice in video object detection [16, 18, 55], the mAP@IoU=0.5 is utilized as an evaluation metric to report results. More details can be found in Appendix A.

**Results.** Table 5 compares our FeatEnHancer with several LLIE methods [44, 21, 14, 67, 17, 25] and the powerful video object detection baseline [55]. Evidently, our FeatEnHancer provides considerable gains to the baseline with 34.6 mAP and 11.2 mAP under illumination levels of 3.2 and 0.2, respectively. Note that our FeatEnHancer is the only method that boosts performance under both image and video modalities. In contrast, as shown in Table 5, existing LLIE methods not only fail to assist the baseline method but also deteriorate the performance. This poor generalization of LLIE approaches highlights that learning from domain-specific paired data [14, 67, 44], unpaired data [21], and curve estimation without data [17, 25] are not the optimal solutions for generalized enhancement methods. Hence, more research is required.

4.5. Ablation Studies

This section ablates important design choices in the proposed FeatEnHancer when plugged into RetinaNet, DeeplabV3+, and SELSA on ExDark (dark object detection), ACDC (nighttime semantic segmentation), and DarkVision with illumination level of 3.2% (video object detection), respectively.

**SAFA in FeatEnHancer.** The important component of the proposed FeatEnHancer is the scale-aware attentional feature aggregation (SAFA) that aggregates high-resolution features. To validate its effectiveness, we conduct multiple experiments where SAFA is replaced with simple averaging or skip connections (SC) [20] to fuse enhanced multi-scale features \( F \) and \( F_q \) (see Sec. 3.2). The experiment results are...
<table>
<thead>
<tr>
<th>Method</th>
<th>ExDark (mAP)</th>
<th>ACDC (mIoU)</th>
<th>DarkVision (mAP)</th>
</tr>
</thead>
<tbody>
<tr>
<td>simple averaging</td>
<td>60.5</td>
<td>50.3</td>
<td>32.9</td>
</tr>
<tr>
<td>skip connections [20]</td>
<td>70.3</td>
<td>51.7</td>
<td>33.1</td>
</tr>
<tr>
<td>SAFA</td>
<td>72.6</td>
<td>54.9</td>
<td>34.6</td>
</tr>
</tbody>
</table>

(a) Effectiveness of SAFA.

<table>
<thead>
<tr>
<th>Method</th>
<th>ExDark (mAP)</th>
<th>ACDC (mIoU)</th>
<th>DarkVision (mAP)</th>
</tr>
</thead>
<tbody>
<tr>
<td>maxpool</td>
<td>69.3</td>
<td>51.3</td>
<td>32.9</td>
</tr>
<tr>
<td>adavgpool [68]</td>
<td>69.9</td>
<td>50.7</td>
<td>32.9</td>
</tr>
<tr>
<td>interpolation [30]</td>
<td>70.7</td>
<td>51.5</td>
<td>33.1</td>
</tr>
<tr>
<td>Convolution</td>
<td>72.6</td>
<td>54.9</td>
<td>34.6</td>
</tr>
</tbody>
</table>

(b) Various combinations of multi-scale fusion.

<table>
<thead>
<tr>
<th>Scale</th>
<th>ACDC (mIoU)</th>
<th>DarkVision (mAP)</th>
</tr>
</thead>
<tbody>
<tr>
<td>(2, 4)</td>
<td>71.8</td>
<td>52.7</td>
</tr>
<tr>
<td>(4, 8)</td>
<td>72.6</td>
<td>54.9</td>
</tr>
<tr>
<td>(8, 16)</td>
<td>68.7</td>
<td>51.4</td>
</tr>
</tbody>
</table>

(c) Downsampling approaches.

Table 6: Ablations for the proposed FeatEnHancer on three benchmarks. (a) We investigate the effectiveness of SAFA by replacing it with different aggregation methods to fuse $F$ and $F_q$. (b) We experiment with various combinations of SAFA and skip connection (SC) to justify an optimal design choice. Here, (SC, SC) means employing only skip connections to merge both $F_q$ and $F_o$ with $F$. (c) Besides convolution, we experiment with other downsampling techniques to generate lower resolutions. Here, adavgpool denotes adaptive average pooling as done in [68]. (d) We vary scale sizes to generate lower-scale representations. Here, $(2, 4)$ means $I_q \in \mathbb{R}^{H/2 \times W/4 \times 3}$ and $I_o \in \mathbb{R}^{H/4 \times W/8 \times 3}$. (e) We vary number of attentional blocks $N$ in SAFA of FeatEnHancer. Default settings are highlighted.

summarized in Table 6a. It is clear that SAFA outperforms both averaging and SC strategies by +2.3 mAP on ExDark, +3.2 mIoU on ACDC, and +1.5 mAP on DarkVision. These significant boosts across all three benchmarks indicate that scale-aware attention leads to optimal multi-scale feature aggregation in the proposed FeatEnHancer.

Multi-scale feature fusion. We experiment with various combinations of SAFA and SC to find an optimal design choice to fuse $F_q$ and $F_o$ with $F$ (see Sec. 3.2). As shown in Table 6b, there is a clear increase in performance, achieving (72.6 mAP on ExDark, 54.9 mIoU on ACDC, and 34.6 mAP on DarkVision) when SAFA is applied to fuse $F$ and $F_q$ first, and then $F_o$ is merged with the output of SAFA using skip connection. Hence, we use this approach as the default setting.

Convolutional Downsampling. Table 6c summarizes results from different downsampling techniques applied on the input image $I$ to generate lower-resolutions $I_q$ and $I_o$ (see Sec 3.1). Our proposed convolutional downsampling yields impressive gains of +1.9 mAP on ExDark, +3.4 mIoU, and +1.5 mAP on DarkVision compared to max-pooling, adaptive average pooling [68], and bilinear interpolation [30]. These results demonstrate the effectiveness of convolutional downsampling since it is better aligned with various vision backbone networks [32, 15, 27].

Different Scale sizes. We analyse the effect of different scale sizes to generate lower resolutions in Table 6d. Here, for instance, $(2, 4)$ means that the resolution of input image $I \in \mathbb{R}^{H \times W \times 3}$ is reduced by a factor of 2 and 4 to generate $I_q \in \mathbb{R}^{H/2 \times W/4 \times 3}$ and $I_o \in \mathbb{R}^{H/4 \times W/8 \times 3}$, respectively. Note that all these scales are generated through regular convolutional operator $\text{Conv}(\cdot)$, as explained in Eq. 1. Looking at results in Table 6d, the top performance on all three tasks is achieved with the scale size of $(4, 8)$, thereby, preferred as a default setting.

Number of Attention Blocks in SAFA. Table 6e studies the effect of the number of attention blocks $N$ in our SAFA. The performance rises for all three tasks with the increase in $N$. This demonstrates that more attentional blocks in SAFA bring additional gains. The best performance with 72.6 mAP on ExDark, 54.9 mIoU on ACDC, and 34.6 mAP on DarkVision is achieved when $N$ reaches 8, and after that, it tends to saturate. Hence, $N = 8$ is used as the default setting.

5. Conclusion

This paper proposes FeatEnHancer, a novel general-purpose feature enhancement module designed to enrich hierarchical features favourable for downstream tasks under low-light vision. Our intra-scale feature enhancement and scale-aware attentional feature aggregation schemes are aligned with vision backbone networks and produce powerful semantic representations. Furthermore, our FeatEnHancer neither requires pre-training on synthetic datasets nor relies on enhancement loss functions. These architectural innovations make FeatEnHancer a plug-and-play module. Extensive experiments on four different downstream vision tasks covering both images and videos demonstrate that our method brings consistent and significant improvements over baselines, LLIE methods, and task-specific state-of-the-art approaches.
References


