Semantic-Guided Zero-Shot Learning for Low-Light Image/Video Enhancement

Shen Zheng
Wenzhou-Kean University
Wenzhou, China
zhengsh@kean.edu

Gaurav Gupta
Wenzhou-Kean University
Wenzhou, China
ggupta@kean.edu

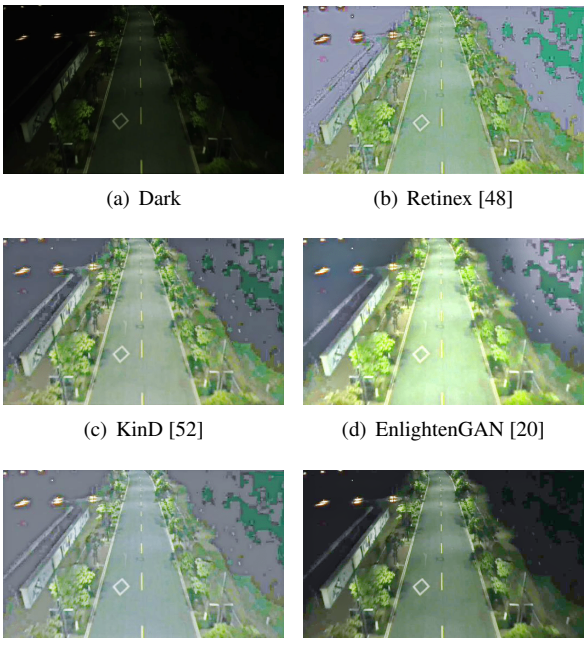
Abstract

Low-light images challenge both human perceptions and computer vision algorithms. It is crucial to make algorithms robust to enlighten low-light images for computational photography and computer vision applications such as real-time detection and segmentation. This paper proposes a semantic-guided zero-shot low-light enhancement network (SGZ) which is trained in the absence of paired images, unpaired datasets, and segmentation annotation. Firstly, we design an enhancement factor extraction network using depthwise separable convolution for an efficient estimate of the pixel-wise light deficiency of an low-light image. Secondly, we propose a recurrent image enhancement network to progressively enhance the low-light image with affordable model size. Finally, we introduce an unsupervised semantic segmentation network for preserving the semantic information during intensive enhancement. Extensive experiments on benchmark datasets and a low-light video demonstrate that our model outperforms the previous state-of-the-art. We further discuss the benefits of the proposed method for low-light detection and segmentation. Code is available at https://github.com/ShenZheng2000/Semantic-Guided-Low-Light-Image-Enhancement.

1. Introduction

[1] Low-light images degraded due to environmental or technical restraints suffer from various problems such as under-exposure and high ISO noise. As a result, those images are prone to have degraded features and contrast, which harm the low-level perceptual quality and deteriorate high-level computer vision tasks relying on accurate semantic information. It is necessary to improve the visual quality and to enhance the generalizability of the advanced vision algorithms.

[1]This work is supported by the research funding from Wenzhou-Kean University with project number SpF2021011. We thank Changjie Lu for his help on model architecture design.



(a) Dark (b) Retinex [48] (c) KinD [52] (d) EnlightenGAN [20] (e) Zero-DCE [11] (f) Ours

Figure 1. The enhancement result on a nighttime aerial video frame. Our proposed model has excellent perceptual quality in terms of exposure, contrast, color, and edge information. In comparison, other models either fail to enhance the dark regions or generate unpleasant noise, blur or artifacts.

One plausible way to increase brightness at low-light conditions is to use higher ISO or more extended exposure time. Nevertheless, those strategies respectively intensify noises and introduce motion blur [2]. The other reasonable approach is to use modern software like Photoshop or Lightroom for light adjustment. However, these software requires artistic skills and are inefficient for large-scale datasets with diverse illumination conditions.

Traditional low-light image enhancement methods mostly involves Histogram Equalization [17, 44] and Retinex theory [24, 47, 8, 9, 13]. Although these methods can generate encouraging perceptual qualities in some

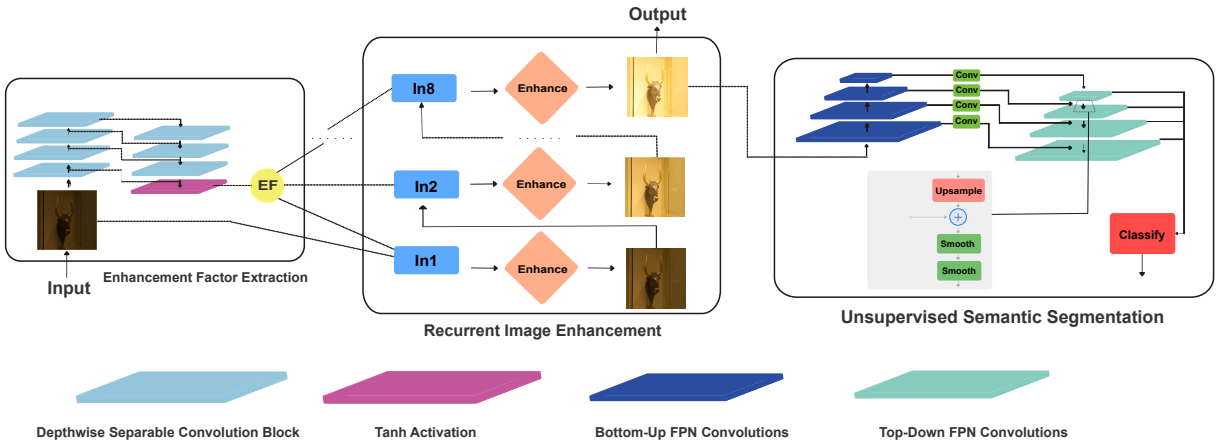


Figure 2. The proposed model architecture. Our model consists of a three stage network: EFE for estimating the light enhancement factor, RIE for progressively lighten the image, and USS for segment the enhanced image. During training, both RIE and USS have frozen parameters and they output the loss to update EFE. During testing, EFE and RIE are used sequentially to enhance an low-light image.

situations, their performances depend on manually selected priors and hand-crafted regularization which are difficult to tune. Furthermore, the long inference time resulting from the intricate optimization process makes them unfitting for real-time tasks.

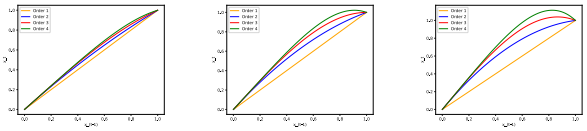
Deep learning based low-light image enhancement methods have recently received much attention due to their compelling efficiency, accuracy, and robustness [26]. Supervised methods [33, 48, 52, 45] have the highest scores at some benchmark datasets [47, 13, 35, 25] with their excellent image-to-image mapping abilities. However, they require paired training images (i.e., low/normal light pairs), which either need expensive retouch or demand unfeasible image capture with the same scene but different lighting conditions. On the other hand, Unsupervised methods [20] require only an unpaired dataset for training. Nonetheless, the data bias from the manually selected datasets restricts their generalization ability. Zero-shot learning [11, 27] methods eliminate the need for both paired images and unpaired dataset. However, they ignore the semantic information, which is shown by [37, 10, 31] to be crucial for high-level vision tasks. As a result, their enhanced images are in sub-optimal visual quality. Fig. 1 reveals the limitations of the previous researches.

To address the limitations discussed above, we present a semantic-guided zero-shot framework for low-light image enhancement (Fig. 2). As we focus on low-light image/video enhancement, we first design a light-weight enhancement factor extraction (EFE) network with depthwise separable convolution [15] and symmetric skip connections. The EFE is highly adaptive and can leverage the spatial information of the low-light images to monitor the subsequent image enhancement. To perform image enhancement with affordable model size, we then introduce a recurrent image enhancement (RIE) network which utilizes both the low-light image and the enhancement factor from EFE as its input. The RIE is able to progressively enhance the images, using the previous stage's output as the input for the subsequent recurrent stage. Aiming to preserve the semantic information during the enhancement process, we finally propose an unsupervised semantic segmentation (USS) network requiring no expensive segmentation annotation. The USS receives the enhanced image from RIE and utilizes feature pyramid network [29] to calculate the segmentation loss. The segmentation loss merges with other non-reference loss functions as the total loss, which updates the parameters of EFE during training.

The contributions of the proposed work are summarized as follows:

- We propose a new semantic-guided zero-shot low-light image enhancement network. To the best of our knowledge, we are the first to fuse high-level semantic information into low-level image enhancement with the absence of paired images, unpaired datasets, or segmentation labels.
- We develop a light-weight convolutional neural network to automatically extract the enhancement factor which record the pixel-wise light deficiency of an low-light image.
- We design an recurrent image enhancement strategy with five non-reference loss functions to boost our model's generalization ability to images of diverse lighting conditions.
- We conduct extensive experiments to demonstrate the superiority of our model in both qualitative and quantitative metrics. **Our model is ideal for low-light video enhancement because it can process 1000 images of size 1200 $\times$ 900 within 1 second on a single GPU**.

Figure 3. Enhancement factor visualization. Left column: Low-Light Images. Right column: Corresponding Enhancement Factor. Darker region indicates lower values for the enhancement factor.



(a) $|x_r| = 0.2$ (b) $|x_r| = 0.5$ (c) $|x_r| = 0.8$

Figure 4. Recurrent image enhancement illustration with different enhancement factor x_r and different Order. The horizontal axis x_{t-1} refers to a pixel's value in the $(t-1)$ stage, whereas the vertical axis x_t refers to that pixel's value in the t stage. All pixel value is scaled to [0, 1]. Greater $|x_r|$ indicates a more intense enhancement.

2. Related Work

Traditional Low-Light Image Enhancement Traditional low-light Image Enhancement mainly consists of histogram equalization (HE-based) methods and Retinex-based methods. HE-based image enhancement methods have been widely applied in the early years. BPDHE [17] proposes a brightness preserving dynamic histogram equalization method that can maintain the mean intensity of the low-light image in its enhanced version. WTHE [44] introduces a contrast enhancement method that performs weighting and thresholding on the histogram of an image before the histogram equalization operation.

Recently, many Retinex-based methods have been designed for low-light image enhancement. NPE [47] proposes a non-uniform, naturalness-preserving enhancement method to balance image details and naturalness. PIE [8] presents a probabilistic enhance approach which exploits concurrent estimation of illumination and reflectance. LIME [13] estimate a coarse illumination map finding the maximum value in the R, G, B channel and then improve that coarse map using a structure prior.

Unlike conventional methods, our model uses a light-weight convolutional neural network to automatically extract the enhancement factor that learns the enlightenment requirement from the low-light images. That design allows the recurrent image enhancement network to run in linear complexity yet still achieving compelling results.

Deep Low-Light Image Enhancement Deep learning based low-light image enhancement methods can be mainly classified into supervised learning, unsupervised learning, and zero-shot learning. The pioneering supervised low-light enhancement method LLNet [33] presents a noise-robust autoencoder-based way to enlighten images with minimum pixel-level saturation. Retinex [48] considers Retinex theory, integrating a decomposition network and an illumination adjustment network that learns from paired low/normal light pictures. The similar work KinD [52] additionally introduce degradation removal in the reflectance.

Unsupervised methods avoid the tedious work for preparing paired training images. EnlightenGAN [20] is the first low-light image enhancement method trained without paired data. It utilizes an attention-based multi-scale discriminator with self-regularized loss functions. Zero-shot learning eliminates the need for both paired images and unpaired datasets. Zero-DCE [11, 27] designs a lightweight network for light-enhancement curves approximation and use non-reference loss functions to enhance the low-light images.

Unlike other deep low-light image enhancement methods, our model exploits the high-level semantic information with a pretrained segmentation network requiring no segmentation label. That design allows us to preserve an essential amount of semantic information without significantly increasing the computational complexity.

3. Proposed Method

3.1. Enhancement Factor Extraction Network

The enhancement factor extraction (EFE) aims to learns the pixel-wise light deficiency of a low-light image and records that information in a enhancement factor. Inspired by the architecture of U-Net [42], EFE is a fully convolutonal neural network with symmetric skips connections, which means that it can address input images of arbitrary size. No batch normalization or up/downsampling is adopted since they will damage the spatial coherence of the enhanced image [43, 21, 18]. Each convolution block in EFE consists of a 3×3 depthwise separable convolution layer and a subsequent ReLU [38] activation layer. The last convolution block reduces the channel numbers from 32 to 3 and output the enhancement factor x_r via the Tanh activation. Fig. 3 visualizes the enhancement factor extracted from 2 low-light images. It is evident that brighter regions in the low-light image corresponds to lower values in the enhancement factor, and vice verse.

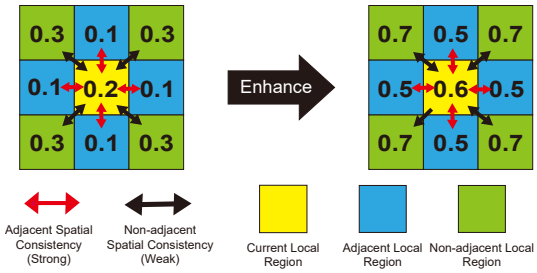


Figure 5. Illustration of spatial consistency loss L_{spa}, assuming the pixel value of the input image is scaled to [0, 1]. L_{spa} encourages the connection between neighboring local regions when the enhancement is increasing all pixel values.

3.2. Recurrent Image Enhancement Network

Inspired by the success of recurrence [41, 55, 28] and light enhancement curve [50, 11] in low-light image enhancement, we build a recurrent image enhancement (RIE) network to enhance the low-light image according to the enhancement factor and then output the enhanced image. Each recurrence considers the previous stage's output and the enhancement factor as its input. The recurrent enhancement process is:

$$x_t = x_{t-1} + x_r * (x_{t-1}^{\text{Order}} - x_{t-1}) \tag{1}$$

where x is the output, x_r is the enhancement factor and t is the recurrence step. The next step is to decide the optimal Order to enlight the image. Since the recurrent network should be simple for differentiation and should be effective for progressive lightening, we only consider positive integers for Order. With this in mind, we plot the recurrent image enhancement with respect to different x_r and Order in Fig. 4. When Order is 1, the pixel value is insensitive to x_r and is the same as the previous stage. When Order equals 3 or 4, the pixel value approaches or even exceed 1.0, making a image looks too bright. In comparison, Order of 2 grants the most robust enhancement in recurrence.

3.3. Unsupervised Semantic Segmentation Network

The Unsupervised Semantic Segmentation (USS) Network aims at accurate pixel-wise segmentation of the enhanced image which preserve the semantic information during progressive image enhancement. Similar to [7, 32, 46, 12], we freeze all the layers for segmentation network during training. Here, we use two pathways, including the bottom-up pathway which uses ResNet-50 [14] with ImageNet [5] weights, and the top-down pathway which uses Gaussian initialization with a mean of 0 and a standard deviation of 0.01. Both pathways have four convolution blocks which is connected to each other through lateral connections. The choice of weight initialization will be explained in the ablation study.

The enhanced image from RIE will first enter the bottom-up pathway for feature extraction. The top-down pathway then transforms the high-semantic layers into high-resolution ones for spatial-aware semantic segmentation. Each convolution block in the top-down approach performs bi-linear upsampling on the image and concatenates it with the lateral outcome. Two smooth layers with 3×3 convolution are applied after the concatenation for better perceptual quality. Finally, we concatenate the result of each block in the top-down pathway and calculate the segmentation.

3.4. Loss Functions

We adopt five non-reference loss functions, including L_{spa}, L_{rgb}, L_{bri}, L_{tv} ,and L_{sem}. We do not consider content loss or perceptual loss [35] due to the unavailability of paired training images.

Spatial Consistency Loss This Spatial Consistency loss helps to maintain the spatial consistency between the low-light image and the enhanced image by conserving the neighbor pixels' differences during enhancement. Unlike [11, 27] that only consider adjacent cells, we also include the spatial coherence with non-adjacent neighbors (See Fig. 5). The spatial consistency loss is:

$$\begin{aligned} L_{spa} = \frac{1}{A}\sum_{i=1}^{A}[\sum_{j\in\phi(i)}(|(Y_i - Y_j)| - |(I_i - I_j)|)^2 + \\ \alpha * \sum_{k\in\psi(i)}(|(Y_i - Y_k)| - |(I_i - I_k)|)^2] \end{aligned} \tag{2}$$

where Y and I are the mean pixel value in a $A \times A$ local region in an enhanced image and the low-light image, respectively. A is the side of the local regions which we set to 4 according to the ablation study. $\phi(i)$ is the four adjacent neighbors (top, down, left, right), and $\psi(i)$ is the four non-adjacent neighbors (top left, top right, lower left, and lower right). α is 0.5 since the weight of the non-adjacent neighbors is less important.

RGB Loss The color loss [45, 52, 11] reduces color incorrectness in the enhanced picture by bridging different color channels. We adopt Charbonnier loss which helps high-quality image reconstruction [23, 19]. The RGB loss is:

$$\begin{aligned} L_{rgb} = \sum_{\forall(i,j)\in\zeta}\sqrt{((Y^i) - (Y^j))^2 + \varepsilon^2}, \\ \zeta = \{(R,G),(R,B),(G,B)\} \end{aligned} \tag{3}$$

where ε is a penalty term that is empirically set to 10^{-6} for training stability.

Brightness Loss Inspired by [34, 45, 11], we design a brightness loss to constrains the under/over-exposure in an image. The loss measures the L1 difference between the

Figure 6. Sample training images. Our training dataset consists of images of different backgrounds and diverse illumination conditions.

average pixel value of a specific region to a predefined exposure level E. The brightness loss is:

$$L_{bri} = \frac{1}{A}\sum_{a=1}^{A}|Y_a - E| \tag{4}$$

where E is the ideal image exposure level which is set to 0.60 according to the ablation study.

Total Variation Loss The total variation loss [3] measures the difference between the neighboring pixels in an image. We use total variation loss here to reduce noise and to increase image smoothness. Unlike prior low-light image enhancement works [48, 52, 45, 11], we additionally consider inter-channel (R, G, and B) relations in the loss to improve the color brightness. Our total variation loss is:

$$L_{tv} = \frac{1}{CHW}\sum_{c=1}^{C}\sum_{h=1}^{H}\sum_{w=1}^{W}\left[(\nabla_x Y_{c,h,w})^2 + (\nabla_y Y_{c,h,w})^2\right] \tag{5}$$

where C, H and W represents channel, height and width of a image, respectively. ∇_x and ∇_y is the horizontal and vertical gradient operations, respectively.

Semantic Loss The semantic loss helps to maintain the semantic information of an image during enhancement. We refer to the focal loss [30] for writing our cost function. Recommended by the ablation study, our semantic loss require no segmentation label and only a pre-initialized model. The semantic loss is:

$$L_{sem} = \frac{1}{HW}\sum_{1\le i\le H, 1\le j\le W} -\beta\,(1-p_{i,j})^{\gamma}\log p_{i,j} \tag{6}$$

where p is the segmentation network's estimated class probability for a pixel. Inspired by [7], we chose the focal coefficient β and γ as 1 and 2, respectively.

Total Loss The total loss function can be summarized as:

$$L_{total} = \lambda_{spa} * L_{spa} + \lambda_{rgb} * L_{rgb} + \lambda_{bri} * L_{bri} + \lambda_{tv} * L_{tv} + \lambda_{sem} * L_{sem} \tag{7}$$

Here, we set $\lambda_{spa} = \lambda_{rgb} = \lambda_{bri} = \lambda_{tv} = 1$ and $\lambda_{sem} = 0.1$.

Name	Number	Format	Type	Metric
NPE[47]	10	RGB	Real	U, B
LIME[13]	84	RGB	Real	U, B
MEF[35]	17	RGB	Real	U, B
DICM[25]	64	RGB	Real	U, B
VV	24	RGB	Real	U, B
LOL[48]	15	RGB	Real	P, S, M
DarkBDD	100	RGB	Real	U, B
DarkCityScape	150	RGB	Synthetic	P, S, M

Table 1. Dataset description. Where U, B stands for UNIQUE and BRISQUE, and P, S, M stands for PSNR, SSIM, MSE, respectively.

4. Experiments

4.1. Implementation Details

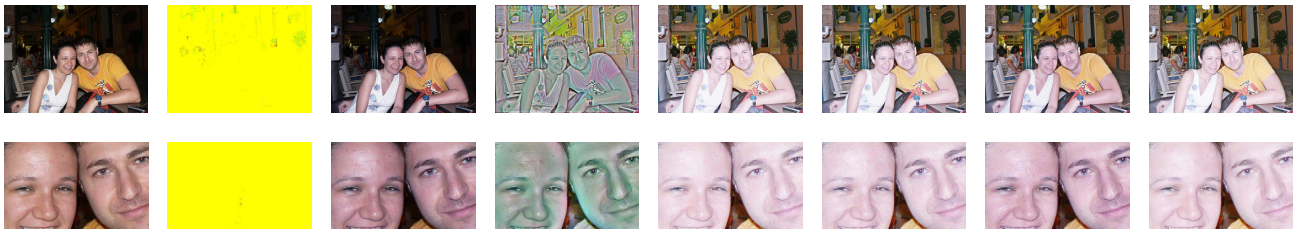
We select 2002 images of different exposure levels and resize them to 512×512 (See Fig. 6) for our model training. The proposed model is trained with Pytorch [39] on a single NVIDIA 2080 Ti GPU for 100 epochs using the Adam [22] optimizer with an initial learning rate of 0.0001. The batch size is 6, which takes around 3 hours to converge. Besides, we clip gradient norm to be within 0.1. For initialization of the EFE, we use a normally distributed weight with zero mean and a standard deviation of 0.02.

4.2. Evaluation Dataset

We consider two traditional methods PIE [8] and LIME [13], three supervised deep learning methods Retinex [48], MBLLEN [34] and KinD [52], one unsupervised method EnlightenGAN [20], and one zero-shot learning method Zero-DCE[11] for model comparison.

Our datasets for comparison includes NPE [47], LIME [13], MEF [35], DICM [25], VV[2] and LOL [48]. Since this paper aims at enhancement of low-light RGB images, we do not include raw datasets such as MIT-Adobe FiveK [1] or SID [2]. Instead, moving towards task-driven low-light image enhancement, we additionally select 100 low-light images from BDD10K [49] and name it DarkBDD. Besides, we use gamma correction on 150 images from the CityScape [4] dataset to synthesize a new dataset called DarkCityScape.

For evaluating the model performance, we use reference metrics including Peak Signal-to-Noise Ratio (PSNR), Structure Similarity Index (SSIM) and Mean Square Error (MSE), and non-reference metric including Unified No-reference Image Quality and Uncertainty Evaluator (UNIQUE) [51] and Blind/Referenceless Image Spatial Quality Evaluator (BRISQUE) [36]. The overall description is in Table 1.



(i) Dark (j) $w/o\ L_{rgb}$ (k) $w/o\ L_{bri}$ (l) $w/o\ L_{tv}$ (m) $w/o\ L_{spa}$ (n) $w/o\ L_{sem}$ (o) $w/o\ L_{spa}\ L_{sem}$ (p) Ours

Figure 7. Visual comparison on loss function ablations. Top row: original enhanced images. Bottom row: cropped enhanced images

Method	PSNR	SSIM	MSE
w/o L_{rgb}	8.840	0.523	0.233
w/o L_{bri}	13.06	0.473	0.108
w/o L_{tv}	16.05	0.678	0.045
w/o L_{spa}	20.53	0.785	**0.023**
w/o L_{sem}	20.15	0.786	0.025
w/o L_{spa} L_{sem}	19.86	0.785	0.026
Ours	**20.60**	**0.793**	**0.023**

Table 2. Ablations of loss function on LOL dataset in terms of PSNR ↑, SSIM ↑ and MSE ↓.

Method	PSNR	SSIM	MSE
A = 3	20.44	0.789	**0.023**
A = 5	20.26	0.787	0.024
E = 0.5	17.25	0.694	0.054
E = 0.7	20.54	0.762	0.024
Ours	**20.60**	**0.793**	**0.023**

Table 3. Ablations of loss function hyperparameters A and E on LOL dataset in terms of PSNR ↑, SSIM ↑ and MSE ↓. The proposed method uses A = 4 and E = 0.6.

4.3. Ablation Study

We conduct ablation studies to investigate individual loss functions, loss function hyperparameters A and E, and weight initialization for USS.

Table 2 displays the loss function ablation. It shows that L_{bri}, L_{rgb} and L_{tv} has large influences on image enhancement results, whereas L_{spa} and L_{sem} have smaller impacts. An additional visual comparison is made on the VV dataset in Fig. 7. We find that Model without L_{rgb}/L_{bri}/L_{tv} have severe color deviation, poor low-light region enhancement, and unnatural artifacts, respectively. We also note that Model without L_{spa} or L_{sem} generates noise-corrupted facial details. Model without L_{spa} and L_{sem} additionally result in poor regional contrast and deficient dark area illumination.

The ablation of loss function hyperparameters is in Table. 3. It can be seen that the proposed value for A and E generates the best result. In short, all loss functions with concurrent settings are essential to reach a promising performance.

[2]https://sites.google.com/site/vonikakis/datasets

Weight	LOL	DarkCityScape
Gaussian	20.60 / 0.79	**25.97 / 0.97**
VOC [6]	**20.63 / 0.81**	24.86 / 0.95

Table 4. Ablations of USS top-down pathway weight initialization on LOL and DarkCityScape dataset in terms of PSNR ↑/SSIM ↑.

The ablation of weight initialization is in Table 4. Although USS pretrained on VOC has slightly better result at LOL, it's performance at the large-scale DarkCityScape is much worse than USS with Gaussian initialization. This phenomenon could result from data bias. Based on this evidence, we conclude that Gaussian initialization is sufficient for a promising outcome.

4.4. Model Comparisons

Quantitative Comparison We conduct quantitative comparisons for different models. Traditional methods PIE [8] and LIME [13] were excluded for efficiency comparison because they unfit GPU acceleration. For all tables, we use **bold** for the best score and blue for the second-best score. '-' indicates that a result is unavailable due to excessive image size for a particular model.

Table 5 shows the comparison on NPE, LIME, MEF, DICM, and VV datasets. Our model has the best average UNIQUE and the second best average BRISQUE. Table 6 shows the model comparison on LOL and DarkCityScape dataset. Our method is the second best in LOL and is the best in the more challenging extreme low-light dataset DarkCityScape.

Table 7 shows that the proposed model is computationally the most efficient. The proposed model's run time is 0.001 second for a single image (i.e., 1000 images can be processed within 1 second). Besides, the significantly fewer FLOPs indicates our model fits low-light video enhancement. Furthermore, the proposed method is ideal for mobile devices due to the small parameters.

Quantitative Comparison We present the visual comparison of different models at Fig. 8. It can be seen that the proposed model significantly enhances the dark regions, main-

Method	NPE[47]	LIME[13]	MEF[35]	DICM[25]	VV	DarkBDD	Average
Dark	0.793 / 19.81	**0.826** / 21.81	0.738 / 23.56	0.795 / **21.57**	0.826 / **23.62**	0.799 / 61.62	0.796 / 28.67
PIE[8]	0.801 / 21.72	0.791 / 22.72	0.752 / **11.02**	0.791 / 21.72	0.832 / 26.54	0.796 / 53.22	0.794 / **26.16**
LIME[13]	0.786 / 18.24	0.774 / 20.44	0.722 / 15.25	0.758 / 23.48	0.820 / 27.14	-/-	-/-
Retinex[48]	**0.828** / 16.04	0.794 / 31.47	0.755 / 20.08	0.770 / 29.53	0.824 / 29.58	0.792 / 50.77	0.794 / 29.57
MBLLEN[34]	0.793 / 34.46	0.768 / 30.26	0.717 / 37.44	0.787 / 32.44	0.719 / 26.13	0.772 / 51.40	0.759 / 35.35
KinD[52]	0.792 / 19.65	0.766 / 39.29	0.747 / 31.36	0.776 / 32.71	0.814 / 29.34	0.778 / **49.38**	0.779 / 33.62
Zero-DCE[11]	0.814 / 17.06	0.811 / 21.40	0.762 / 16.84	0.777 / 27.35	0.835 / 24.26	0.800 / 59.37	0.800 / 27.71
Ours	0.786 / **13.25**	0.807 / **19.99**	**0.785** / 13.92	**0.801** / 26.12	**0.836** / 31.72	**0.815** / 57.06	**0.805** / 27.01

Table 5. UNIQUE ↑ / BRISQUE ↓ Comparison on NPE, LIME, MEF, DICM, VV and DarkBDD

Dataset	Dark	PIE[8]	Retinex[48]	MBLLEN[34]	KinD[52]	Zero-DCE[11]	**Ours**
LOL	13.20/0.48/0.106	20.18/0.77/0.025	17.59/0.54/0.044	**21.21/0.84/0.016**	19.29/0.76/0.040	20.38/0.78/0.023	20.60/0.79/0.023
DCS	16.22/0.77/0.026	17.49/0.83/0.020	10.54/0.65/0.091	22.52/0.88/0.007	12.28/0.73/0.062	22.59/0.94/0.006	**25.97/0.97/0.004**

Table 6. PSNR ↑ / SSIM ↑ / MSE ↓ Comparison on LOL and DarkCityScape (DCS)

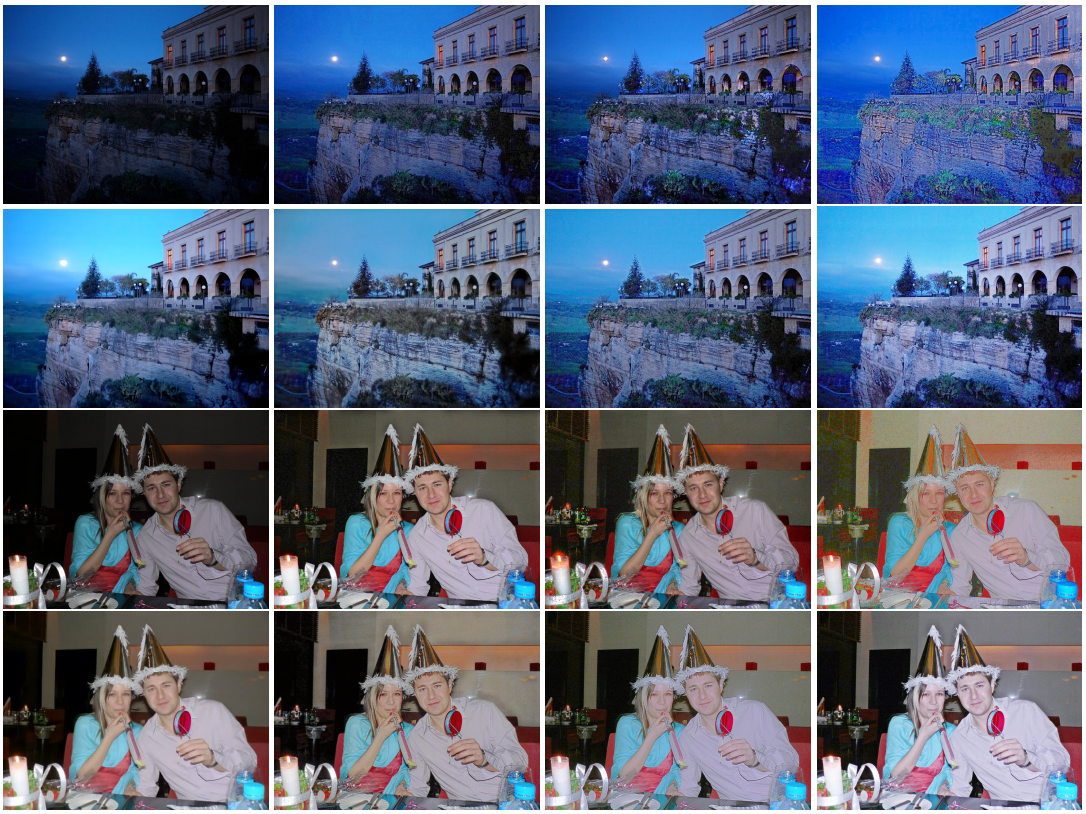


Figure 8. Visual Comparison on LIME [13] (top two rows) and VV dataset (bottom two rows). For each two rows, from left to right, and from top to bottom: Dark, PIE[8], LIME[13], Retinex[48], MBLLEN[34], KinD[52], Zero-DCE[11], Ours

Figure 9. Object Detection Results on DarkBDD. From left to right, and from top to bottom: Dark, PIE[8], Retinex[48], MBLLEN[34], KinD[52], EnlightenGAN[20], Zero-DCE[11], Ours

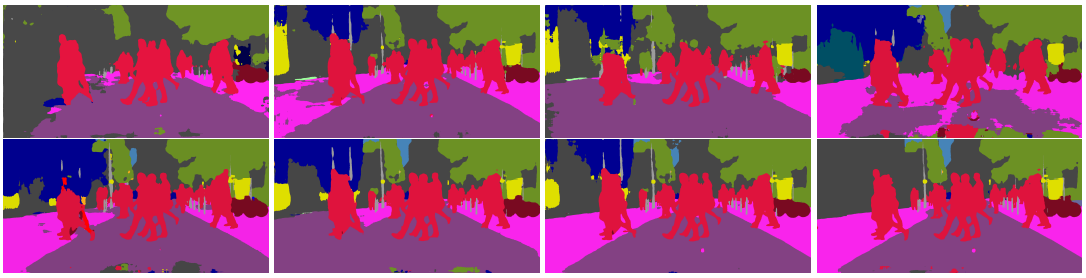


Figure 10. Semantic Segmentation Results on DarkCityScape. From left to right, and from top to bottom: Dark, PIE[8], Retinex[48], MBLLEN[34], KinD[52], Zero-DCE[11], Ours, GroundTruth

Method	RT↓	Params↓	FLOPs↓	Score↑
Retinex[48]	0.121	0.555	587.5	2.30
MBLLEN[34]	0.526	0.450	301.1	3.05
KinD[52]	0.147	8.160	575.0	3.36
EnlightenGAN[20]	0.008	8.637	273.2	2.94
Zero-DCE[11]	0.003	0.079	84.99	2.60
Ours	**0.001**	**0.011**	**0.120**	**4.04**

Table 7. Model Efficiency and User Study Score Comparison. Images of size 1200×900 are selected for experiments. 'RT' is the inference time in seconds per image. 'Params' are the numbers of trainable parameters in millions per image, and 'FLOPs' are the numbers of floating-point operations in billions per image. All model inference is conducted with a single Nvidia GeForce RTX 2080 Ti GPU.

tains color balance and image contrast, and presents natural exposure with sufficient facial detail.

4.5. Low-Light Detection and Segmentation

We utilize the object detection model Yolov3 [40] and the semantic segmentation model PSPNet [53] to investigate how different low-light image enhancement methods are beneficial to the high-level tasks.

We show the perceptual comparison of object detection in Fig. 9. PIE, Retinex, and Zero-DCE improve the image's brightness but meanwhile introduce blur and noise. KinD and EnlightenGAN, though somewhat aids detection, produces unnatural background artifacts. In comparison, our model helps detect the greatest numbers of cars.

We display the perceptual comparison of semantic segmentation in Fig. 10. Retinex and MBLLEN leave large areas of incorrect segmentation. PIE, Zero-DCE and KinD's enhancement leads to accurate pedestrians segmentation but undesirable holes in the left sidewalk and background trees. In comparison, the proposed method is the closest to the groundtruth. Finally, we show a quantitative comparison on semantic segmentation using mean Intersection Over Union (mIOU) and mean Pixel Accuracy (mPA) in Table 8. Our model has the best score for both mIOU and mPA.

Metric	Dark	PIE	Retinex	MBLLEN	KinD	Zero-DCE	**Ours**
mIOU	54.49	61.97	57.96	51.98	63.42	64.36	**65.87**
mPA	70.76	68.89	66.76	59.06	71.69	74.20	**74.50**

Table 8. mIOU (%) ↑ and mPA (%) ↑ Comparison on DarkCityScape

4.6. Low-Light Video Enhancement

Unlike prior researches that have been pivotal to single low-light image enhancement, we also examine the performance on a nighttime aerial video. The video is captured using a drone camera with 24 FPS and a resolution of 960×540. The video is 41 seconds long and is saved as MP4. We conduct a user study to quantitatively assess the enhancement performance. Specifically, we ask 50 adult participants to rate the enhancement result (video) of five models, including EnlightenGAN [20], KinD [52], Retinex [48], Zero-DCE [11] and our model, and we report the result in Table 7. More video enhancement results are in the supplementary material.

Figure 11. The failure cases of the proposed model. Left two: low-light images. Right two: our enhanced results. Our model cannot address strong motion blurs or mirror reflection.

5. Conclusion

This paper introduced a novel semantic-guided zero-shot low-light image enhancement network. The proposed network is trainable without paired images, unpaired datasets, or segmentation labels. That is achieved by enhancement factor extraction, recurrent image enhancement, and unsupervised semantic segmentation. Extensive experiments demonstrated the excellence of the proposed method in terms of perceptual quality, model efficiency, and the benefits for high-level vision tasks. Our future plan is to investigate motion blur removal with low-light image/video enhancement [16]. We also intend to explore detection-driven enhancement algorithms [54].

References

[1] Vladimir Bychkovsky, Sylvain Paris, Eric Chan, and Frédo Durand. Learning photographic global tonal adjustment with a database of input/output image pairs. In *CVPR 2011*, pages 97–104. IEEE, 2011.

[2] Chen Chen, Qifeng Chen, Jia Xu, and Vladlen Koltun. Learning to see in the dark. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3291–3300, 2018.

[3] Qiang Chen, Philippe Montesinos, Quan Sen Sun, Peng Ann Heng, et al. Adaptive total variation denoising based on difference curvature. *Image and vision computing*, 28(3):298–306, 2010.

[4] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3213–3223, 2016.

[5] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[6] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88(2):303–338, 2010.

[7] Zhiwen Fan, Liyan Sun, Xinghao Ding, Yue Huang, Congbo Cai, and John Paisley. A segmentation-aware deep fusion network for compressed sensing mri. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 55–70, 2018.

[8] Xueyang Fu, Yinghao Liao, Delu Zeng, Yue Huang, Xiao-Ping Zhang, and Xinghao Ding. A probabilistic method for image enhancement with simultaneous illumination and reflectance estimation. *IEEE Transactions on Image Processing*, 24(12):4965–4977, 2015.

[9] Xueyang Fu, Delu Zeng, Yue Huang, Xiao-Ping Zhang, and Xinghao Ding. A weighted variational model for simultaneous reflectance and illumination estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[10] Abel Gonzalez-Garcia, Davide Modolo, and Vittorio Ferrari. Objects as context for detecting their semantic parts. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6907–6916, 2018.

[11] Chunle Guo, Chongyi Li, Jichang Guo, Chen Change Loy, Junhui Hou, Sam Kwong, and Runmin Cong. Zero-reference deep curve estimation for low-light image enhancement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1780–1789, 2020.

[12] Mengxi Guo, Mingtao Chen, Cong Ma, Yuan Li, Xianfeng Li, and Xiaodong Xie. High-level task-driven single image deraining: Segmentation in rainy days. In *International Conference on Neural Information Processing*, pages 350–362. Springer, 2020.

[13] Xiaojie Guo, Yu Li, and Haibin Ling. Lime: Low-light image enhancement via illumination map estimation. *IEEE Transactions on image processing*, 26(2):982–993, 2016.

[14] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[15] Andrew G Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861*, 2017.

[16] Zhe Hu, Sunghyun Cho, Jue Wang, and Ming-Hsuan Yang. Deblurring low-light images with light streaks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3382–3389, 2014.

[17] Haidi Ibrahim and Nicholas Sia Pik Kong. Brightness preserving dynamic histogram equalization for image contrast enhancement. *IEEE Transactions on Consumer Electronics*, 53(4):1752–1758, 2007.

[18] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1125–1134, 2017.

[19] Kui Jiang, Zhongyuan Wang, Peng Yi, Chen Chen, Baojin Huang, Yimin Luo, Jiayi Ma, and Junjun Jiang. Multi-scale progressive fusion network for single image deraining. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8346–8355, 2020.

[20] 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.

[21] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *European conference on computer vision*, pages 694–711. Springer, 2016.

[22] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[23] Wei-Sheng Lai, Jia-Bin Huang, Narendra Ahuja, and Ming-Hsuan Yang. Fast and accurate image super-resolution with deep laplacian pyramid networks. *IEEE transactions on pattern analysis and machine intelligence*, 41(11):2599–2613, 2018.

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

[25] Chulwoo Lee, Chul Lee, and Chang-Su Kim. Contrast enhancement based on layered difference representation. In *2012 19th IEEE International Conference on Image Processing*, pages 965–968. IEEE, 2012.

[26] Chongyi Li, Chunle Guo, Linghao Han, Jun Jiang, Ming-Ming Cheng, Jinwei Gu, and Chen Change Loy. Lighting the darkness in the deep learning era. *arXiv preprint arXiv:2104.10729*, 2021.

[27] Chongyi Li, Chunle Guo, and Chen Change Loy. Learning to enhance low-light image via zero-reference deep curve estimation. *arXiv preprint arXiv:2103.00860*, 2021.

[28] Jinjiang Li, Xiaomei Feng, and Zhen Hua. Low-light image enhancement via progressive-recursive network. *IEEE Transactions on Circuits and Systems for Video Technology*, 2021.

[29] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2117–2125, 2017.

[30] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017.

[31] Daqi Liu, Miroslaw Bober, and Josef Kittler. Visual semantic information pursuit: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 2019.

[32] Ding Liu, Bihan Wen, Xianming Liu, Zhangyang Wang, and Thomas S Huang. When image denoising meets high-level vision tasks: A deep learning approach. *arXiv preprint arXiv:1706.04284*, 2017.

[33] Kin Gwn Lore, Adedotun Akintayo, and Soumik Sarkar. Llnet: A deep autoencoder approach to natural low-light image enhancement. *Pattern Recognition*, 61:650–662, 2017.

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

[35] Kede Ma, Kai Zeng, and Zhou Wang. Perceptual quality assessment for multi-exposure image fusion. *IEEE Transactions on Image Processing*, 24(11):3345–3356, 2015.

[36] Anish Mittal, Anush Krishna Moorthy, and Alan Conrad Bovik. No-reference image quality assessment in the spatial domain. *IEEE Transactions on image processing*, 21(12):4695–4708, 2012.

[37] Roozbeh Mottaghi, Xianjie Chen, Xiaobai Liu, Nam-Gyu Cho, Seong-Whan Lee, Sanja Fidler, Raquel Urtasun, and Alan Yuille. The role of context for object detection and semantic segmentation in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 891–898, 2014.

[38] Vinod Nair and Geoffrey E Hinton. Rectified linear units improve restricted boltzmann machines. In *Icml*, 2010.

[39] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *arXiv preprint arXiv:1912.01703*, 2019.

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

[41] Wenqi Ren, Sifei Liu, Lin Ma, Qianqian Xu, Xiangyu Xu, Xiaochun Cao, Junping Du, and Ming-Hsuan Yang. Low-light image enhancement via a deep hybrid network. *IEEE Transactions on Image Processing*, 28(9):4364–4375, 2019.

[42] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.

[43] Dmitry Ulyanov, Andrea Vedaldi, and Victor Lempitsky. Instance normalization: The missing ingredient for fast stylization. *arXiv preprint arXiv:1607.08022*, 2016.

[44] Qing Wang and Rabab K Ward. Fast image/video contrast enhancement based on weighted thresholded histogram equalization. *IEEE transactions on Consumer Electronics*, 53(2):757–764, 2007.

[45] Ruixing Wang, Qing Zhang, Chi-Wing Fu, Xiaoyong Shen, Wei-Shi Zheng, and Jiaya Jia. Underexposed photo enhancement using deep illumination estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6849–6857, 2019.

[46] Sicheng Wang, Bihan Wen, Junru Wu, Dacheng Tao, and Zhangyang Wang. Segmentation-aware image denoising without knowing true segmentation. *arXiv preprint arXiv:1905.08965*, 2019.

[47] Shuhang Wang, Jin Zheng, Hai-Miao Hu, and Bo Li. Naturalness preserved enhancement algorithm for non-uniform illumination images. *IEEE Transactions on Image Processing*, 22(9):3538–3548, 2013.

[48] Chen Wei, Wenjing Wang, Wenhan Yang, and Jiaying Liu. Deep retinex decomposition for low-light enhancement. *arXiv preprint arXiv:1808.04560*, 2018.

[49] Fisher Yu, Wenqi Xian, Yingying Chen, Fangchen Liu, Mike Liao, Vashisht Madhavan, and Trevor Darrell. Bdd100k: A diverse driving video database with scalable annotation tooling. *arXiv preprint arXiv:1805.04687*, 2(5):6, 2018.

[50] Lu Yuan and Jian Sun. Automatic exposure correction of consumer photographs. In *European Conference on Computer Vision*, pages 771–785. Springer, 2012.

[51] Weixia Zhang, Kede Ma, Guangtao Zhai, and Xiaokang Yang. Uncertainty-aware blind image quality assessment in the laboratory and wild. *IEEE Transactions on Image Processing*, 30:3474–3486, 2021.

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

[53] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2881–2890, 2017.

[54] Shen Zheng, Yuxiong Wu, Shiyu Jiang, Changjie Lu, and Gaurav Gupta. Deblur-yolo: Real-time object detection with efficient blind motion deblurring. In *2021 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE, 2021.

[55] Minfeng Zhu, Pingbo Pan, Wei Chen, and Yi Yang. Eemefn: Low-light image enhancement via edge-enhanced multi-exposure fusion network. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 13106–13113, 2020.