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
gupta@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 a 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

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.

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
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.
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. WTIE [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 lightweight 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.


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 convolutional 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 $3 \times 3$ 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 versa.
Adjacent Spatial Consistency (Strong)
Non-adjacent Spatial Consistency (Weak)
Current Local Region
Adjacent Local Region
Non-adjacent Local Region

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:

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

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 $3 \times 3$ 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 $L_{spa}$, $L_{rgb}$, $L_{bru}$, $L_{tue}$, 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:

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

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). $\alpha$ 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:

$$L_{rgb} = \sum_{\forall (i, j) \in \zeta} \sqrt{(Y_i - Y_j)^2 + \epsilon^2},$$

where $\zeta = \{(R, G), (R, B), (G, B)\}$ and $\epsilon$ 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
average pixel value of a specific region to a predefined exposure level $E$. The brightness loss is:

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

(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:

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

(5)

where $C$, $H$ and $W$ represents channel, height and width of a image, respectively. $\nabla_x$ and $\nabla_y$ 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:

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

(6)

where $p$ is the segmentation network’s estimated class probability for a pixel. Inspired by [7], we chose the focal coefficient $\beta$ and $\gamma$ as 1 and 2, respectively.

**Total Loss** The total loss function can be summarized as:

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

(7)

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

---

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

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

4. Experiments

4.1. Implementation Details

We select 2002 images of different exposure levels and resize them to $512 \times 512$ (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

- 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.

- Our datasets for comparison includes NPE [47], LIME [13], MEF [35], DICM [25], VV2 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.
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 $L_{tv}$ have 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.

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-
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Dark</td>
<td>0.793</td>
<td>0.826</td>
<td>0.738</td>
<td>0.795</td>
<td>0.215</td>
<td>0.826</td>
<td>0.799</td>
</tr>
<tr>
<td>PIE[8]</td>
<td>0.801</td>
<td>0.791</td>
<td>0.752</td>
<td>0.791</td>
<td>0.217</td>
<td>0.832</td>
<td>0.796</td>
</tr>
<tr>
<td>LIME[13]</td>
<td>0.786</td>
<td>0.774</td>
<td>0.722</td>
<td>0.758</td>
<td>0.234</td>
<td>0.820</td>
<td>0.794</td>
</tr>
<tr>
<td>Retinex[48]</td>
<td>0.828</td>
<td>0.794</td>
<td>0.755</td>
<td>0.770</td>
<td>0.293</td>
<td>0.824</td>
<td>0.792</td>
</tr>
<tr>
<td>MBLLEN[34]</td>
<td>0.793</td>
<td>0.768</td>
<td>0.717</td>
<td>0.787</td>
<td>0.244</td>
<td>0.719</td>
<td>0.772</td>
</tr>
<tr>
<td>KinD[52]</td>
<td>0.792</td>
<td>0.766</td>
<td>0.747</td>
<td>0.776</td>
<td>0.271</td>
<td>0.814</td>
<td>0.778</td>
</tr>
<tr>
<td>Zero-DCE[11]</td>
<td>0.814</td>
<td>0.811</td>
<td>0.762</td>
<td>0.777</td>
<td>0.273</td>
<td>0.835</td>
<td>0.840</td>
</tr>
<tr>
<td><strong>Ours</strong></td>
<td>0.807</td>
<td>0.807</td>
<td>0.807</td>
<td>0.807</td>
<td>0.807</td>
<td>0.807</td>
<td>0.807</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>LOL</td>
<td>13.20/0.48/0.106</td>
<td>20.18/0.77/0.025</td>
<td>17.59/0.54/0.044</td>
<td><strong>21.21/0.84/0.016</strong></td>
<td>19.29/0.76/0.040</td>
<td>20.38/0.78/0.023</td>
<td>20.60/0.79/0.023</td>
</tr>
<tr>
<td>DCS</td>
<td>16.22/0.77/0.026</td>
<td>17.49/0.83/0.020</td>
<td>10.54/0.65/0.091</td>
<td>22.52/0.88/0.007</td>
<td>22.52/0.73/0.062</td>
<td>22.59/0.94/0.006</td>
<td>25.97/0.97/0.004</td>
</tr>
</tbody>
</table>

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
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.

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


