PSENet: Progressive Self-Enhancement Network for Unsupervised Extreme-Light Image Enhancement

Hue Nguyen   Diep Tran   Khoi Nguyen   Rang Nguyen
VinAI Research, Vietnam
{v.huent88, v.dieptn147, v.khoindm, v.rangnhm}@vinai.io

Abstract

The extremes of lighting (e.g. too much or too little light) usually cause many troubles for machine and human vision. Many recent works have mainly focused on under-exposure cases where images are often captured in low-light conditions (e.g. nighttime) and achieved promising results for enhancing the quality of images. However, they are inferior to handling images under over-exposure. To mitigate this limitation, we propose a novel unsupervised enhancement framework which is robust against various lighting conditions while does not require any well-exposed images to serve as the ground-truths. Our main concept is to construct pseudo-ground-truth images synthesized from multiple source images that simulate all potential exposure scenarios to train the enhancement network. Our extensive experiments show that the proposed approach consistently outperforms the current state-of-the-art unsupervised counterparts in several public datasets in terms of both quantitative metrics and qualitative results. Our code is available at https://github.com/VinAIResearch/PSENet-Image-Enhancement.

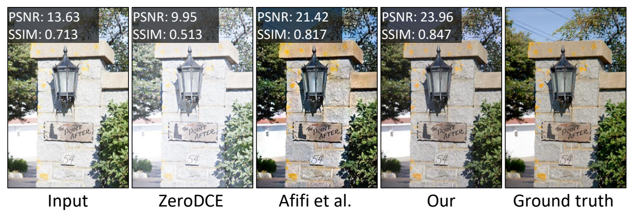

Figure 1. Visual comparison on an over-exposed scene. Most of the previous state-of-the-art failed to recover the over-exposed case except recent work proposed by Afifi *et al.* [1], which is trained with full supervision.

1. Introduction

Producing images with high contrast, vivid color, and rich details is one of the important goals of photography. However, acquiring such pleasing images is not always a trivial task due to harsh lighting conditions, including extreme low lighting or unbalanced lighting conditions caused by backlighting. The resulting under-/over-exposed images usually decrease not only human satisfaction but also computer vision system performance on several downstream tasks such as object detection [31] or image segmentation [35]. Wrong exposure problems occur early in the capturing process and are difficult to fix once the final 8-bit image has been rendered. This is because the in-camera image signal processors usually use highly nonlinear operations to generate the final 8-bit standard RGB image [26, 12, 27].

Many recent works have mainly focused on under-exposure cases where images are often captured in low-light conditions (*e.g.* nighttime). These works have achieved promising results for enhancing the quality of images even captured under extreme low-light conditions. However, they failed to handle over-exposure images, as shown in Fig. 1. The recent work proposed by Afifi *et al.* [1] achieves impressive results in improving both under- and over-exposed cases. However, their proposed method is designed to work in a supervised manner, requiring a large dataset of wrongly exposed and corresponding ground-truth (GT) well-exposed image pairs. This data collection is typically time-consuming and expensive.

In this paper, we propose a novel unsupervised approach that does not require any well-exposed GT images. The key idea is to generate a pseudo GT image given the input wrongly exposed one in order to train an enhancement network. The pseudo GT image across training epochs is progressively generated by choosing the visually best regions taken from multiple sources, namely the output of the same input image from the previous epoch, the brighter/darker reference images by changing the gamma value of the input image, and the input image itself. The choosing criteria are well-exposedness, local contrast, and color saturation, which are driven by human knowledge of a visually good image and have been shown to be effective in measuring perceptual image quality [23]. In this way, the task of generating pseudo GT images is simply comparing and selecting the best regions from different sources where almost possible cases of exposure are simulated in training. Furthermore, by using the output of the previous epoch as a source for choosing, we ensure that the output of the current epoch will be better than or at least equal to that of the previous one, giving the name of our approach PSENet – Progressive Self Enhancement Network.

Our contributions are summarized as follows:

- We introduce a new method for generating effective pseudo-GT images from given wrongly-exposed images. The generating process is driven by a new non-reference score reflecting the human evaluation of a visually good image.
- We propose a novel unsupervised progressive pseudo-GT-based approach that is robust to various severe lighting conditions, i.e. under-exposure and over-exposure. As a result, the burden of gathering the matched image pairs is removed.
- Comprehensive experiments are conducted to show that our approach outperforms previous unsupervised methods by large margins on the SICE [3] and Afifi [1] datasets and obtains comparable results with supervised counterparts.

2. Related Work

Image enhancement approaches can be divided into two categories: traditional and learning-based methods.

Traditional methods. One of the simplest and fastest approaches is to transform single pixels of an input image by a mathematical function such as linear function, gamma function, or logarithmic function. For example, histogram equalization-based algorithms stretch out the image's intensity range using the cumulative distribution function, resulting in the image's increased global contrast. The Retinex theory [16], on the other hand, argues that an image is made from two components: reflectance and illumination. By estimating the illumination component of an image, the dynamic range of the image can be easily adjusted to reproduce images with better color contrast. However, most Retinex algorithms use Gaussian convolution to estimate illumination, thus leading to blurring edges [38]. Frequency-domain-based methods, by contrast, preserve edges by employing the high-pass filter to enhance the high-frequency components in the Fourier transform domain [36]. However, the adaptability of such traditional methods is often limited due to their unawareness of the overall and local complex gray distribution of an image [38]. For a systematic review of conventional approaches, we suggest readers refer to the work of Wang *et al*. [38].

Learning-based methods. In recent years, there has been increasing attention to learning-based photo-enhancing methods in both supervised and unsupervised manners.

Supervised learning methods aim to recover natural images by either directly outputting the high quality images [21, 22, 19, 40] or learning specific parameters of a parametric model (*e.g*. Retinex model) [6, 37, 24] from a paired dataset. SID [5] is a typical example in the first direction. In this work, the authors collect a short-exposure low-light image dataset and adopt a vanilla Unet architecture [32] to produce an enhanced sRGB image from raw data thus replacing the traditional image processing pipeline. Following this work, Lamba and Mitra [14] present a novel network architecture that concurrently processes all the scales of an image and can reduce the latency times by 30% without decreasing the image quality. Different from the previously mentioned approaches, Cai *et al*. [3] explore a new direction in which both under and over-exposed images are considered. They introduce a novel two-stage framework trained on their own multi-exposure image dataset, which enhances the low-frequency and high-frequency components separately before refining the whole image in the second stage. Afifi *et al*. [1] put a further step in this direction by introducing a larger dataset along with a coarse-to-fine neural network to enhance image qualities in both under- and over-exposure cases. For learning a parametric model, Retinex theory [15] is often adopted [39, 18, 37]. Benefiting from paired data, the authors focus on designing networks to estimate the reflectance and illumination of an input image. Dealing with the image enhancement task differently, HDRNet [6] presents a novel convolutional neural network to predict the coefficients of a locally-affine model in bilateral space using pairs of input/output images.

Unsupervised learning. Collecting paired training data is always time-consuming and expensive. To address this issue, an unpaired GAN-based method named EnlightenGAN is proposed in [10]. The network, including an attention-guided U-Net as a generator and global-local discriminators, shows promising results even though the corresponding ground truth image is absent. To further reduce the cost of collecting reference ground truth images, a set of methods [42, 44, 8, 17] that do not require paired or unpaired training data are proposed. Two recent methods in this category named ZeroDCE [8] and Zheng and Gupta [43] show impressive results in low-light image enhancement tasks by using a CNN model trained under a set of no reference loss functions to learn an image-specific curve for producing a high-quality output image. However, these methods seem to perform poorly when extending to correct over-exposed images, as shown in Fig. 1.

Our proposed method, in contrast, is the first deep learning work handling these extreme lighting conditions in an unsupervised manner.

3. Methodology

Given an sRGB image, I, captured under a harsh lighting condition with low contrast and washed-out color, our

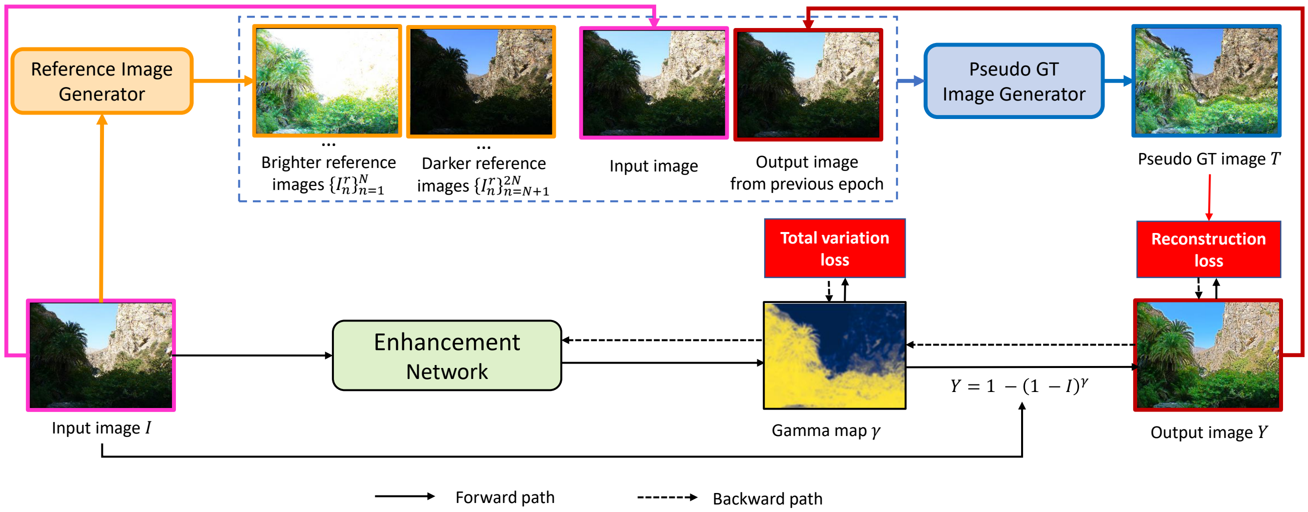

Figure 2. Overview of our proposed framework which comprises three main modules: reference image generator, pseudo GT image generator, and enhancement network. Given the input image I, the reference image generator randomly generates multiple reference images with different exposure values, half of them are brighter than the input image while the rest are darker. The pseudo GT image generator then takes the input image, the output image from the previous epoch, and the generated reference images as input to produce the pseudo GT image T which is visually better than each of the input components alone based on our proposed non-reference scoring criteria. Finally, the enhancement network predicts the gamma map γ to transform the original image I to obtain the output image Y. The enhancement network is trained with two loss functions: reconstruction loss between the output image Y and the pseudo GT image T, and the total variation loss applied on the predicted gamma map γ to imply the smoothness in prediction. It is worth noting that only the enhancement network is used in testing.

approach aims to reconstruct the corresponding enhanced image Y, which is visibly better and visually pleasant in terms of contrast and color without any supervision.

To address the problem, our key contribution is to propose a new self-supervised learning strategy for training the image enhancement network. That is, we randomly synthesize a set of reference images to be combined together to produce a synthetically high-quality GT image for training. The way of combination is driven by the human knowledge of how visually good an image is. To our best knowledge, our *unsupervised* method is the first to produce pseudo-GT images for training on a large set of ill-exposed images; while other data synthesized methods use well-exposed images as GT to generate corresponding ill-exposed inputs. By using this approach, our model does not suffer from the domain gap issue. Compared with image fusion, which only produces a single output image for input, our pseudo GT images are progressively improved after each epoch, making our model adapt to a wide range of lighting conditions (see Sec. 4 for empirical evidence).

In detail, our reference image generator first takes an image as input and generates $2N$ images where the first N images are darker and the rest are brighter compared to the original input image. Then, the pseudo GT generator module uses these reference images along with the input and the previous prediction of the enhancement network to create the pseudo GT image. It is worth noting that including the previous prediction in the set of references ensures that the quality of the pseudo GT image is greater or at least equal to the previous prediction according to our proposed non-reference score, thus making our training progressively improved. Our training framework is illustrated in Fig. 2 and the detail of each module will be described in the following sections.

3.1. Random Reference Image Generation

To synthesize an under/over-exposed image, we employ a gamma mapping function, which is a nonlinear operation often used to adjust the overall brightness of an image in the image processing pipeline [29]. The gamma mapping function is based on the observation that human eyes perceive the relative change in the light following a power-law function rather than a linear function as in cameras [13]. The connection between the gamma mapping function and the human visual system enables the gamma mapping function to be widely used in image contrast enhancement [7, 30, 34]. However, rather than apply the gamma function directly to the original image, we adopt a haze removal technique in which we apply it to the inverted image to generate $2N$ reference images Y_n as shown in Eq. (1). The reason is that hazy images and poor lighting images normally share the same property of low dynamic range with the high noise level. Therefore, haze removal techniques (e.g. using an inverse image) can be used to enhance

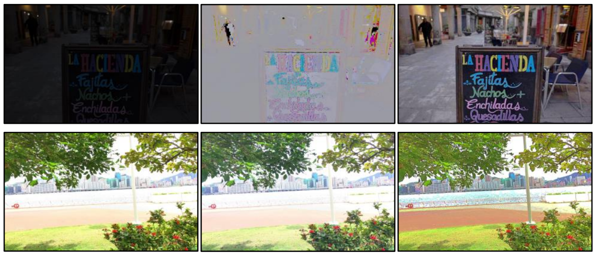

Figure 3. Outputs of the gamma and invert gamma mappings. The output of the latter is visibly better than that of the former.

poor lighting images. When negative images are employed, we discovered that the contrast of images may be improved easier, thus producing a more visually pleasant image, as shown in Fig. 3. In addition, our proposed function also has the same form as the well-known mathematical image processing model LIP [11], which had been proven with both physical and human visual systems, thus making our mapping function more theoretically sound.

$$Y_n = 1 - (1 - I)^{\gamma_n}, \tag{1}$$

where γ_n is a random number whose logarithm value $X_n = \log(\gamma_n)$ is sampled as follows: $X_n \sim U(0, 3)$ for under-exposed reference images and $X_n \sim U(-2, 0)$ for over-exposed reference images.

3.2. Pseudo Ground-truth Image Generator

To create the pseudo GT image, we compare and combine the $2N$ generated reference images, the original image, and the output of the enhancement network of the same image in the previous epoch. The idea behind our approach is inspired by prior work on exposure fusion [23] where a set of perceptual quality measurements for each pixel are calculated on the reference image sequence. These measurements can encode desired attributes such as brightness, contrast, and saturation and have shown their effectiveness in generating a high-quality high dynamic range (HDR) image from an exposure sequence. Therefore, in this paper, we adopt the high-level concept of these measurements but propose a new formulation for each term and a new way to combine these terms together to produce the pseudo GT image.

Well-exposedness of a pixel estimates how likely a pixel belongs to a well-exposed region. We use the L1 distance between the average intensity value of a local region to the well-exposed level, which is set to 0.5. Thus, the well-exposedness of a pixel is defined as:

$$E(x, y) = \left|\mu_{p_{xy}} - 0.5\right|, \tag{2}$$

where p_{xy} is a patch $K \times K$ centered at (x, y) and $\mu_{p_{xy}}$ is its mean intensity value. We set $K = 25$ in this paper.

Local contrast is the local variance of all pixels $I(u, v)$ of the patch $p_{x,y}$

$$C(x, y) = \frac{1}{K \times K} \sum_{u,v \in p_{x,y}} \left[I(u, v) - \mu_{p_{xy}}\right]^2. \tag{3}$$

Color saturation of a pixel measures its vividness. We use the saturation channel in the HSV color space to measure the color saturation of a pixel. In this color space, the saturation is defined as:

$$S(x, y) = \frac{\max(R, G, B) - \min(R, G, B)}{\max(R, G, B)}, \tag{4}$$

where R, G, B correspond to red, green, and blue values of the pixel (x, y).

Since a good-looking output image is one with a low well-exposedness value but high contrast and saturation values, we obtain the final pseudo GT image T by selecting the best regions from all reference images Y_n as follows:

$$T(x, y) = Y_n(x, y) \quad \text{with } n = \operatorname*{argmax}_{n=1,\ldots,2N+2} \frac{C_n(x, y) S_n(x, y)}{E_n(x, y)}. \tag{5}$$

It is worth noting that the way we use the final score is completely different from that of [23]. In [23], the final score map is used as the weight in the weighted sum to combine the reference images together in order to obtain the pseudo GT image. In contrast, we use the final score map as an image comparison tool to select the best regions from the reference images.

3.3. Image Enhancement Network

Similar to [8], our network will learn to predict the intermediate parameters of a tone mapping function instead of directly predicting the output image. We design our lightweight enhancement network based on the UNet architecture [32].

The detail of network architecture is provided in the *Supp. material*. For consistency with the reference image generation module, the invert gamma mapping function in Sec. 3.1 is used to produce the final image Y given the predicted gamma map γ and the original image I:

$$Y = 1 - (1 - I)^{\gamma}. \tag{6}$$

We then train our model end-to-end to minimize the following loss function:

$$L = L_{rec} + \alpha L_{tv}, \tag{7}$$

where L_{rec} and L_{tv} are the reconstruction loss and total variation loss, respectively, and α is a coefficient to balance between the two losses.

Reconstruction loss. We adopt the mean squared error between the network prediction and the pseudo GT image as follows:

$$L_{rec} = \frac{1}{3HW} \sum_{c,x,y} \left[\hat{Y}(c,x,y) - T(c,x,y) \right]^2, \quad (8)$$

where c is the color channel, $\hat{Y}$ is the output image and T is our generated pseudo GT image in Sec. 3.2; H and W are the height and width of the input image, respectively.

Total variation loss. In the homogeneous areas, the adjacent gamma values should be similar to avoid sudden changes, which can create visual artifacts. Therefore, we apply a familiar smoothness prior to image restoration tasks, called total variation minimization [33, 25], to the predicted gamma map. The total variation loss is defined in Eq. (9)

$$\mathcal{L}_{tv} = \frac{1}{3HW} \sum_{c,x,y} \{ |\gamma_c(x+1,y) - \gamma_c(x,y)| + |\gamma_c(x,y+1) - \gamma_c(x,y)| \}, \quad (9)$$

where γ_c is the predicted gamma value corresponding to the color channel c.

4. Experiments

Datasets. We assess our approach as well as comparative methods on two main multi-exposure datasets: Afifi (introduced by Afifi *et al*. [1]) and SICE [3] datasets. The Afifi dataset contains 24,330 sRGB images rendered from the MIT-Adobe FiveK dataset [2] by varying their digital exposure settings. The SICE dataset has two parts 1 and 2 with 360 and 229 multi-exposure sequences (sets of images of the same scene captured at different exposure levels), respectively. We employ part 1 as the training set and part 2 as the test set. For generalization evaluation, we also test all approaches on the LOL dataset [39], which is composed of 500 pairs of low-light and normal images.

Implementation details. We train our image enhancement network on an NVIDIA A100 GPU, using the Adam optimizer with a batch size of 64. In the SICE dataset, our model is trained with 140 epochs while the number of epochs for training the Afifi dataset is 30. The learning rate is $5e^{-4}$ and is reduced by half on a plateau with the patience of 5. All the input images are resized to 256×256 during training. The coefficient of the total variation loss α for training on each dataset and other implementation details are empirically selected and reported in the *Supp. material*.

4.1. Comparison with Prior Work

We compare our approach with two traditional methods: CLAHE [28], IAGCWD [4], one unpaired method EnlightenGAN [10], and two unsupervised methods: ZeroDCE [8], Zheng and Gupta [43]. We also include the two supervised methods: HDRNet [6], Afifi *et al*. [1] for the reference purpose only. The results of these methods are reproduced by using their public source codes with the recommended parameters.

Objective image quality assessment. We adopt two standard referenced metrics: peak signal-to-noise ratio (PSNR) and structural similarity index measure (SSIM) in the cases where the ground-truth images are available in the testing set (for evaluation purposes only). To see the effectiveness of our method on each type of lighting condition, we also report the metric scores on lighting-dependent subsets of the Afifi dataset [1], which includes 3,543 over-exposed images and 2,362 under-exposed images.

Table 1 reports the quantitative results obtained by each method on the Afifi, SICE, and LOL datasets. On the Afifi dataset, our approach outperforms all state-of-the-art unpaired model EnlightenGAN [10] and unsupervised models ZeroDCE [8], and ZeroDCE++ [17] with significant margins (+5 and +0.1 in PSNR and SSIM metrics, respectively). On the SICE dataset, our approach also surpasses all other unsupervised methods on the SICE dataset with large margins (+3 and +0.02 in PSNR and SSIM metrics, respectively). When compared with the supervised methods, surprisingly, the proposed method obtains better results than HDRNet [6] and Afifi *et al*. [1] in the SSIM index. We further assess the generalization abilities of all methods on the LOL dataset. In this experiment, we report the results of all methods trained on the SICE dataset without further tuning. As can be seen in Tab. 1, the same trend can be observed in which we outperform all unsupervised and unpaired approaches and perform slightly worse than a supervised model HDRNet [6]. However, HDRNet [6] tends to produce output images with visual artifacts that are shown in the next section.

Subjective image quality assessment. A visual comparison among unsupervised approaches with typical underexposure and over-exposure scenes is presented in Fig. 4. Our model is the only one that works on both under- and over-exposure. In the case of underexposure, CLAHE [28] and EnlightenGAN [10] could not brighten the image to a proper level, while ZeroDCE [8] tends to produce an image with washed-out colors. Regarding the over-exposure situation, only CLAHE seems to produce a decent output image whereas ZeroDCE and EnlightenGAN appear to fail to recover the image's details in such a condition. A visual comparison with other supervised methods is also shown in Fig. 5. As mentioned previously, although HDRNet [6] achieved the best PSNR and SSIM scores in most of the cases, however, it often produces output images with visible artifacts (shown in Fig. 5).

User study. For a more convincing evaluation, we also conduct a user study with 260 participants on 100 scenes from

Method	Afifi						SICE		LOL	
	Under		Over		Full					
	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
HDRNet (S)	19.35	0.817	20.65	0.846	20.13	0.834	17.25	0.683	17.29	0.766
Afifi *et al.* (S)	18.88	0.845	19.05	0.850	18.98	0.848	-	-	-	-
CLAHE (N)	16.67	0.780	18.19	0.806	17.58	0.796	13.89	0.610	10.02	0.427
IAGCWD (N)	13.23	0.681	17.99	0.820	16.08	0.765	14.09	0.635	11.26	0.519
ZeroDCE (U)	15.36	0.783	11.85	0.739	13.25	0.757	14.28	0.657	14.16	0.654
Zheng and Gupta (U)	16.69	0.806	11.58	0.726	13.62	0.758	12.54	0.626	14.89	0.675
EnlightenGAN (U*)	14.28	0.752	14.05	0.766	14.14	0.762	14.60	0.680	12.32	0.596
PSENet (U)	**18.82**	**0.858**	**19.72**	**0.875**	**19.36**	**0.869**	**17.74**	**0.704**	**16.60**	**0.693**

Table 1. Results on SICE, Afifi, and LOL dataset. The higher the better. The best results are in bold. The terms "under" and "over" stand for under-exposure and over-exposure subsets. The terms "N", "U", "U*" "S" stand for non-learning, unsupervised, unpaired, and supervised, respectively. All methods are trained with the same training sets but different supervision levels. Due to the Matlab license issue, we could not train and evaluate the performance of Afifi *et al*. on the SICE and the LOL datasets. Note that HDRNet [6] and Afifi et al.'s method [1] are both supervised methods (faded in gray, solely for reference purpose).

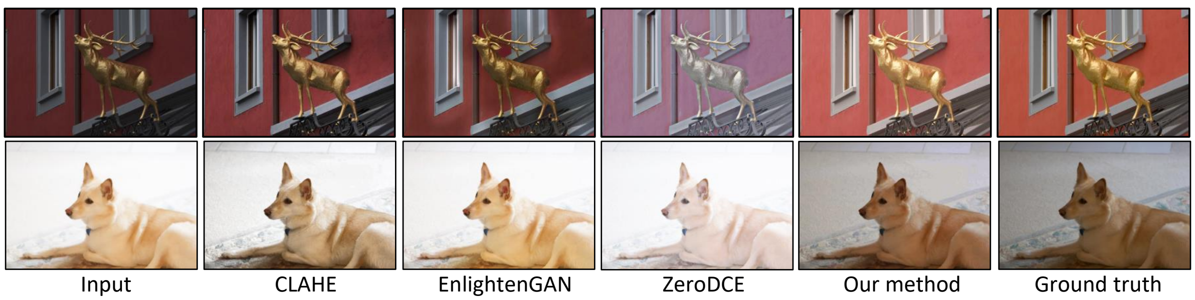

Figure 4. Visual comparison with unsupervised and traditional methods. In the under-exposure situation, CLAHE [28] and Enlighten-GAN [10] could not brighten the image to a proper level, while ZeroDCE [8] tends to produce an image with washed-out colors. As for the over-exposed image, only CLAHE seems to produce an acceptable output whereas ZeroDCE and EnlightenGAN appear to fail to recover the image's details in such a condition.

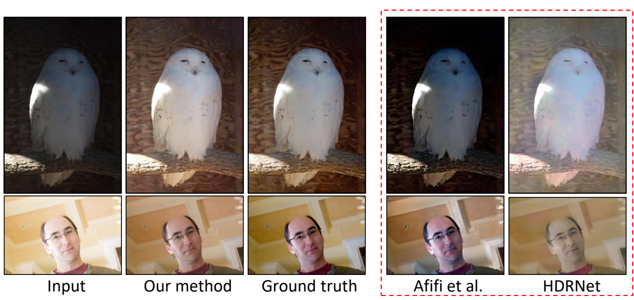

Figure 5. Visual comparison with supervised methods (put in red box). Although HDRNet [6] achieved the best PSNR and SSIM scores in most of the cases, it often produces output images with visible artifacts.

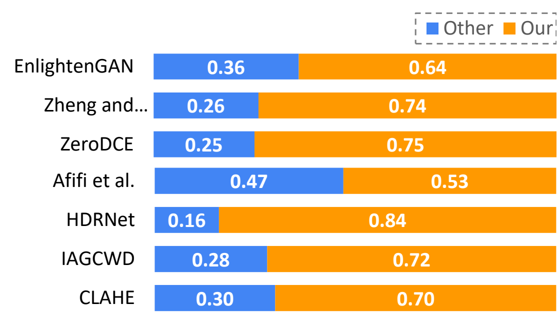

Figure 6. Results of user study - Others vs Ours. The blue color shows the preference percentages of the other methods, while the orange color shows ours.

the testing set to assess human preference for the enhanced results. Out of 100 scenes, 30 scenes are randomly selected to show for each participant, and for each scene, our enhanced image along with another image, which is a result of a random method, is presented. We believe that showing the results of two methods at a time is more reliable than showing the results of all eight methods and asking the users to either rank all methods or choose the best one only. The former is error-prone since the users need to rank pairwise $C(8, 2) = 28$ times for each question. The latter is not informative if ours is not the best. We restrict the sampling method to ensure that all the methods appear evenly across user responses. The participants then are asked to pick a better image in each pair based on the three following criteria: (1) whether all parts in the image are clearly visible; (2) whether the result introduces any color deviation; and (3) the better image based on their preference. The detailed

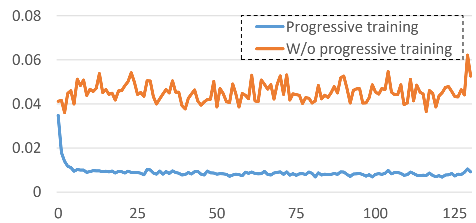

Figure 7. Mean squared error (MSE) between pseudo GT images at two consecutive epochs with and without PTS.

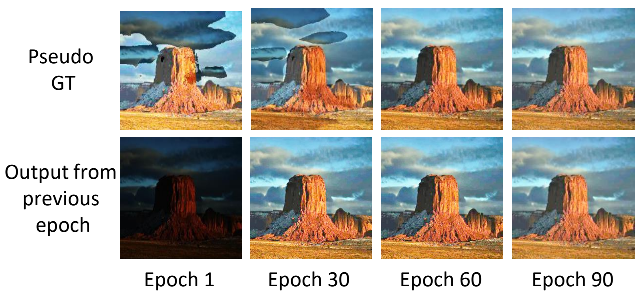

Figure 8. Examples of the pseudo GT image and model output from previous epoch while training. Note that, for epoch 1, the output in the previous epoch is equivalent to the input image. For the first few epochs, the pseudo GT is unstable due to its dependence on the randomness of the reference image generator. However, as the quality of the model's output increases and surpasses reference images, the pseudo GT image tends to converge thanks to the PTS. In addition, the output of our model in these epochs is very close to the pseudo-GTs except for some hash regions, providing useful attention for our model at the latter training stage.

Method	PSNR
ZeroDCE	14.28
ZeroDCE + our pseudo GT	15.30
EnlightenGAN	14.60
EnlightenGAN + our pseudo GT	15.34

Table 2. The influence of our pseudo GT generator on ZeroDCE [8] and EnlightenGAN [10] on the SICE dataset. Our suggested approach also improves these two networks' abilities in handling over-exposure cases, resulting in a considerable improvement (+1 in PSNR).

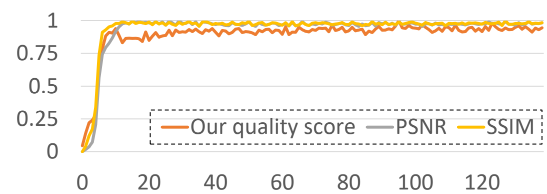

Figure 9. Average PSNR, SSIM (compared with provided GT images), and our proposed image quality score of our pseudo-GT over training epochs (x-axis).

Figure 10. The influence of the total variation loss. Without the loss, the gamma values of neighboring regions are not so smooth, thus breaking down the image structure.

comparison between ours and other methods is reported in Fig. 6. As can be seen, our enhanced images are preferred in all cases including both supervised and unsupervised approaches, only Afifi et al.'s approach [1] has a preferable ratio that is relatively comparable to ours. HDRNet [6] is less desirable in our user study due to their unpleasant output images.

4.2. Ablation Study

In this section, we conduct experiments on analyzing the stability of our method and the impact of different components of our proposed framework. Other experiments related to hyper-parameter selection are presented in the *Supp. material.*

Training stability. Since our method relies on a random reference image generator to produce pseudo GT images, a concern might be raised that whether or not this random factor affects our model's performance. To answer this question, we have retrained our model 10 times with different random seeds. The average performance on SICE is 17.69 ± 0.11 in PSNR and 0.704 ± 0.0013 in SSIM showing that our method is stable regardless of random sampling. This stability is achieved through the progressive training strategy (PTS). In Fig. 7, with the PTS, the MSE between two consecutive pseudo GT decreases when the number of training epochs increases, suggesting that the PTS ensures training convergence and stability. This assumption are also demonstrated in Fig. 8. As a result, our model's final performance also gets better with approximately $+1.5$ and $+0.02$ in terms of PNSR and SSIM, respectively in comparison with the normal training method.

The influence of pseudo GT generator in other enhancement networks. We analyze the influence of our pseudo GT generator by replacing our Enhancement Network with the network presented in ZeroDCE [8] and EnlightenGAN [10], and retrain these models on the SICE dataset. As demonstrated in Tab. 2, using our proposed generator significantly improves the overall performance of these two networks on the SICE dataset (+1 in PSNR). Qualitative results are provided in the *Supp. material*.

The correlation between our proposed quality score and the similarity between the pseudo GT images and reference GT images. We conduct additional experiments to evaluate the correlation between our proposed quality score and the similarity between pseudo GT images and reference GT images measured by PSNR and SSIM metrics. The results in Fig. 9 show that our quality score is an effective measurement of image quality without GT images.

Method	RT (ms)	#Params	#GMACs
EnlightenGAN	94.38	8,636,675	197.11
ZeroDCE	37.87	79,416	61.59
Zheng and Gupta	36.86	10,561	7.834
Our method	20.08	15,251	1.804

Table 3. Running time (**RT**), # of parameters (**#Params**), and # of multiply–accumulated operations (**#GMACs**).

Contribution of total variation loss. We also present the results of our enhancement network trained with the absence of total variation loss in Fig. 10. Without this loss, our model tends to break the relation between neighboring regions, thus breaking down the image structure.

4.3. Computational cost comparison

We evaluate the computational cost of our model and other methods and report the results in Tab. 3. The runtime is measured on on Tesla T4 GPU by processing 50 images of size 1080×720. The number of multiply–accumulated operations (MACs) and the number of trainable parameters for each network are also presented. As we can see, our method is the fastest and extremely lightweight, making it very suitable for real-time applications.

5. Application

In this section, we conduct experiments to evaluate the usefulness of our approach on the face detection task for both under- and over-exposure cases. To the best of our knowledge, there is no public face dataset that contains sufficient samples from both under- and over-exposure for validation. Therefore, we synthetically create a new face dataset from the FDDB dataset [9] by generating new images with different gamma values. The Dual Shot Face Detector (DSFD) [20] trained on the WIDER FACE dataset [41] is used as a pre-trained face detector. More concretely, we feed the images enhanced by several different image enhancement methods to the pre-trained face detector and observe its performance changes.

Fig. 11 depicts the true-positive rate when the number of false-positive samples equals 500, which is computed by the evaluation tool provided in the FDDB dataset [9]. As can be seen, with our image enhancer, DSFD [20] achieves better metric scores consistently on both too dark (low gamma value) and too bright (high gamma value) images. Meanwhile, other methods such as ZeroDCE [8], Zheng and Gupta [43] and EnlightenGAN [10] perform poorly in the over-exposure cases, resulting in a decrease in face detection performance. This demonstrates the robustness of our method under various lighting conditions.

We also present the output of DFSD on two real images where our model is utilized as a preprocessing module in Fig. 12. As can be seen, our model can recover the image detail in both extremely dark or over-bright regions, thus improving the performance of the face detector.

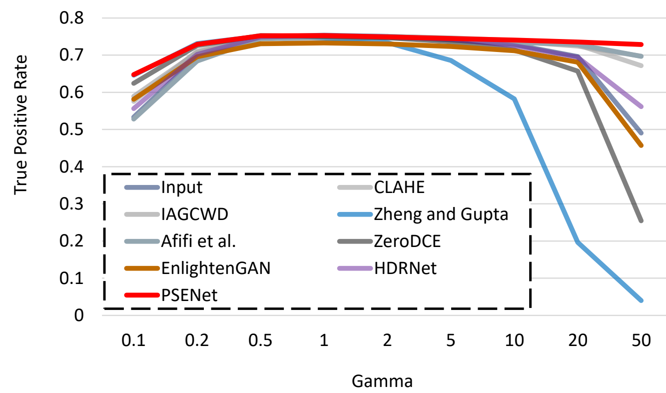

Figure 11. True positive rate when # of false-positive samples equals 500 for different gamma values on the FDDB dataset [9].

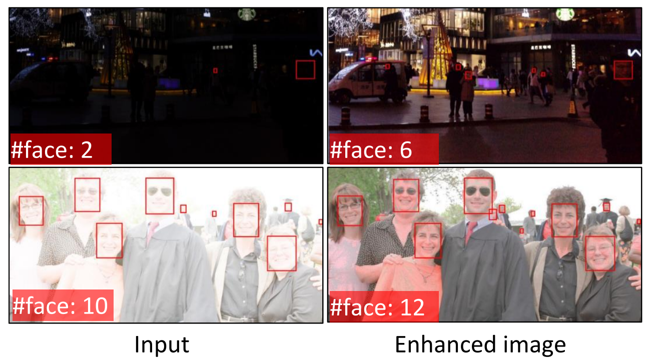

Figure 12. Outputs of the detection model on the images before and after preprocessed by our method. #face denotes the number of faces that are successfully detected by the DFSD.

6. Conclusion

We have introduced a novel progressive self-enhancement network PSENet for image enhancement that is robust to a variety of severe lighting conditions, including under-exposure and over-exposure. In particular, we have developed a new method for generating effective pseudo GT images for training our extreme-light enhancement network in an unsupervised manner. As a result, the burden of gathering the matched photographs is removed. Our extensive experiments show that the proposed approach consistently outperforms previous unsupervised methods by large margins on several public datasets and obtains comparable results with supervised counterparts. We also demonstrate the superior performance of PSENet over all other approaches in the application of face detection in both under-exposure and over-exposure settings. These results justify the importance of PSENet not only in pleasing human vision but also in improving machine perception.

References

[1] Mahmoud Afifi, Konstantinos G. Derpanis, Björn Ommer, and Michael S. Brown. Learning multi-scale photo exposure correction. In *CVPR*, 2021.

[2] 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.

[3] Jianrui Cai, Shuhang Gu, and Lei Zhang. Learning a deep single image contrast enhancer from multi-exposure images. *TIP*, 2018.

[4] Gang Cao, Lihui Huang, Huawei Tian, Xianglin Huang, Yongbin Wang, and Ruicong Zhi. Contrast enhancement of brightness-distorted images by improved adaptive gamma correction. *Computers & Electrical Engineering*, 2018.

[5] Chen Chen, Qifeng Chen, Jia Xu, and Vladlen Koltun. Learning to see in the dark. In *CVPR*, 2018.

[6] Michaël Gharbi, Jiawen Chen, Jonathan T Barron, Samuel W Hasinoff, and Frédo Durand. Deep bilateral learning for real-time image enhancement. *TOG*, 2017.

[7] Xu Guan, Su Jian, Pan Hongda, Zhang Zhiguo, and Gong Haibin. An image enhancement method based on gamma correction. In *International Symposium on Computational Intelligence and Design*, 2009.

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

[9] Vidit Jain and Erik Learned-Miller. Fddb: A benchmark for face detection in unconstrained settings. Technical report, University of Massachusetts, Amherst, 2010.

[10] 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. *TIP*, 2021.

[11] M. Jourlin and J. Pinoli. Logarithmic image processing: The mathematical and physical framework for the representation and processing of transmitted images. *Advances in Imaging and Electron Physics*, 2001.

[12] Hakki Can Karaimer and Michael S Brown. A software platform for manipulating the camera imaging pipeline. In *ECCV*, 2016.

[13] Lester E Krueger. Reconciling fechner and stevens: Toward a unified psychophysical law. *Behavioral and Brain Sciences*, 1989.

[14] Mohit Lamba and Kaushik Mitra. Restoring extremely dark images in real time. In *CVPR*, 2021.

[15] E. Land and J. McCann. Lightness and retinex theory. *JOSA*, 1971.

[16] Edwin H. Land and John J. McCann. Lightness and retinex theory. *JOSA*, 1971.

[17] Chongyi Li, Chunle Guo Guo, and Chen Change Loy. Learning to enhance low-light image via zero-reference deep curve estimation. In *TPAMI*, 2021.

[18] Chongyi Li, Jichang Guo, F. Porikli, and Yanwei Pang. Lightennet: A convolutional neural network for weakly illuminated image enhancement. *Pattern Recognition Letters*, 2018.

[19] Jiaqian Li, Juncheng Li, Faming Fang, Fang Li, and Guixu Zhang. Luminance-aware pyramid network for low-light image enhancement. *Transactions on Multimedia*, 2020.

[20] Jian Li, Yabiao Wang, Changan Wang, Ying Tai, Jianjun Qian, Jian Yang, Chengjie Wang, Jilin Li, and Feiyue Huang. Dsfd: Dual shot face detector. In *CVPR*, 2019.

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

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

[23] Tom Mertens, Jan Kautz, and Frank Van Reeth. Exposure fusion. In *Pacific Conference on Computer Graphics and Applications*, 2007.

[24] Sean Moran, Pierre Marza, Steven McDonagh, Sarah Parisot, and Gregory G. Slabaugh. Deeplpf: Deep local parametric filters for image enhancement. In *CVPR*, 2020.

[25] Michael K Ng and Wei Wang. A total variation model for retinex. *Journal on Imaging Sciences*, 2011.

[26] Rang MH Nguyen and Michael S Brown. Raw image reconstruction using a self-contained srgb-jpeg image with only 64 kb overhead. In *CVPR*, 2016.

[27] Rang MH Nguyen and Michael S Brown. Raw image reconstruction using a self-contained srgb–jpeg image with small memory overhead. *IJCV*, 2018.

[28] Stephen M Pizer, R Eugene Johnston, James P Ericksen, Bonnie C Yankaskas, and Keith E Muller. Contrast-limited adaptive histogram equalization: speed and effectiveness. In *Conference on Visualization in Biomedical Computing*, 1990.

[29] Charles Poynton. *Digital video and HD: Algorithms and Interfaces*. Elsevier, 2012.

[30] Shanto Rahman, Md Mostafijur Rahman, Mohammad Abdullah-Al-Wadud, Golam Dastegir Al-Quaderi, and Mohammad Shoyaib. An adaptive gamma correction for image enhancement. *Journal on Image and Video Processing*, 2016.

[31] Hazem Rashed, Mohamed Ramzy, Victor Vaquero, Ahmad El Sallab, Ganesh Sistu, and Senthil Yogamani. Fusemodnet: Real-time camera and lidar based moving object detection for robust low-light autonomous driving. *ICCVW*, 2019.

[32] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *MICCAI*, pages 234–241, 2015.

[33] L. Rudin, S. Osher, and E. Fatemi. Nonlinear total variation based noise removal algorithms. *Physica D: Nonlinear Phenomena*, 1992.

[34] H. Singh, N. Agrawal, A. Kumar, G. K. Singh, and H.-N. Lee. A novel gamma correction approach using optimally clipped sub-equalization for dark image enhancement. In *International Conference on Digital Signal Processing*, 2016.

[35] Xin Tan, Ke Xu, Ying Cao, Yiheng Zhang, Lizhuang Ma, and Rynson WH Lau. Night-time scene parsing with a large real dataset. *TIP*, 2021.

[36] Liviu I. Voicu, Harley R. Myler, and Arthur Robert Weeks. Practical considerations on color image enhancement using homomorphic filtering. *Journal of Electronic Imaging*, 1997.

[37] Ruixing Wang, Qing Zhang, Chi-Wing Fu, Xiaoyong Shen, Wei-Shi Zheng, and Jiaya Jia. Underexposed photo enhancement using deep illumination estimation. In *CVPR*, 2019.

[38] Wencheng Wang, Xiaojin Wu, Xiaohui Yuan, and Zairui Gao. An experiment-based review of low-light image enhancement methods. *IEEE Access*, 2020.

[39] Chen Wei, Wenjing Wang, Wenhan Yang, and Jiaying Liu. Deep retinex decomposition for low-light enhancement. In *BMVC*, 2018.

[40] Ke Xu, X. Yang, Baocai Yin, and Rynson W. H. Lau. Learning to restore low-light images via decomposition-and-enhancement. *CVPR*, 2020.

[41] Shuo Yang, Ping Luo, Chen Change Loy, and Xiaoou Tang. Wider face: A face detection benchmark. In *CVPR*, 2016.

[42] L. Zhang, Lijun Zhang, Xinyu Liu, Ying Shen, Shaoming Zhang, and Shengjie Zhao. Zero-shot restoration of back-lit images using deep internal learning. *ACM MM*, 2019.

[43] Shen Zheng and Gaurav Gupta. Semantic-guided zero-shot learning for low-light image/video enhancement. In *WACVW*, 2022.

[44] Anqi Zhu, L. Zhang, Ying Shen, Yong Ma, Shengjie Zhao, and Yicong Zhou. Zero-shot restoration of underexposed images via robust retinex decomposition. *ICME*, 2020.