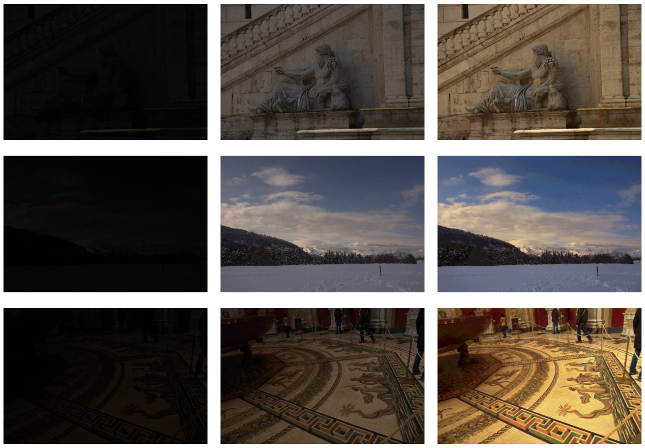
A Two-stage Unsupervised Approach for Low Light Image Enhancement

Junjie Hu, Xiyue Guo, Junfeng Chen, Guanqi Liang, Fuqin Deng, and Tin Lun Lam

***Abstract*—As vision based perception methods are usually built on the normal light assumption, there will be a serious safety issue when deploying them into low light environments. Recently, deep learning based methods have been proposed to enhance low light images by penalizing the pixel-wise loss of low light and normal light images. However, most of them suffer from the following problems: 1) the need of pairs of low light and normal light images for training, 2) the poor performance for dark images, 3) the amplification of noise. To alleviate these problems, in this paper, we propose a two-stage unsupervised method that decomposes the low light image enhancement into a pre-enhancement and a post-refinement problem. In the first stage, we pre-enhance a low light image with a conventional Retinex based method. In the second stage, we use a refinement network learned with adversarial training for further improvement of the image quality. The experimental results show that our method outperforms previous methods on four benchmark datasets. In addition, we show that our method can significantly improve feature points matching and simultaneous localization and mapping in low light conditions.**

***Index Terms*—Low light image enhancement, unsupervised method, SLAM, robot's perception**



(a) Low light inputs. (b) Ground truth. (c) Our results.

Fig. 1. Results of enhancement for three dark images. Our method demonstrates superior performance of enhancement for dark images, as seen that the perceptual quality after enhancement is even better than the ground truth.

I. INTRODUCTION

In recent years, vision based algorithms have brought significant progresses for robot's perception on various tasks such as simultaneous localization and mapping (SLAM) [2], object recognition [10], depth estimation [12], [18], and semantic segmentation [23], [21], [20], etc. However, these algorithms are built upon the assumption that images are captured in a good illumination condition. It captures a serious concern when deploying them into real-world low light environments. As known that low light images especially dark images suffer from poor visibility and high noise, and thus only a little or non-useful information can be used to perform high level perception from them even using powerful deep neural networks. Therefore, it's necessary to enhance low light images in advance.

Recently, deep learning based methods have been continuously proposed to enhance low light images. These methods learn a convolutional network with paired low light and corresponding normal light images in a supervised fashion. Although we have seen great progress made by them, there are mainly three problems that hinder the real-world deployment of those learning based methods. 1) First, it's a challenge to simultaneously acquire low light images from real-world scenes with their corresponding normal light images. Alternatively, researchers have introduced the use of synthesized low light images, however, the model learned from them cannot be directly deployed into real-world scenarios due to domain shift. 2) Second, it's difficult to deal with extremely low light conditions. Deep learning based methods have demonstrated satisfactory performance for slightly low light images, however, they don't perform well for dark images. 3) Besides, low light images usually suffer from strong noise due to the low signal-to-noise ratio, this also brings a difficulty when enhancing low illumination images.

Most of the previous studies for low light image enhancement are focused on handling one of the above problems. For the first problem, researchers have begun to propose unsupervised low light image enhancement approaches. Jiang et al. proposed EnlightenGAN [14] that enhances low light images with a generative adversarial network. Zhang et al. [35] proposed a self-supervised learning based method that can complete the training with even one single low light image based on maximum entropy. For the second problem, Chen et al.[4] propose to recover normal images from extremely dark images by learning a convolutional network with raw

Manuscript received: October, 15, 2020; Accepted December, 13, 2020.

This paper was recommended for publication by Editor Cesar Cadena Lerma upon evaluation of the Associate Editor and Reviewers' comments. This work was supported by the National Natural Science Foundation of China (62073274), and the funding 2019-INT008 from the Shenzhen Institute of Artificial Intelligence and Robotics for Society. (*Corresponding author: Tin Lun Lam.)*

Authors are with the Shenzhen Institute of Artificial Intelligence and Robotics for Society (AIRS), The Chinese University of Hong Kong, Shenzhen, Guangdong, China. (e-mail: hujunjie@cuhk.edu.cn; guoxiyue@cuhk.edu.cn; chenjunfeng@cuhk.edu.cn; guanqiliang@link.cuhk.edu.cn; dengfuqin@cuhk.edu.cn; tllam@cuhk.edu.cn)

Digital Object Identifier (DOI): see top of this page.

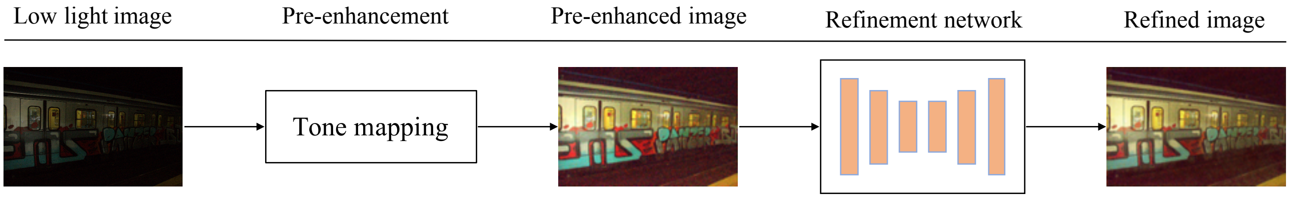


Fig. 2. Diagram of the proposed two-stage framework for low light image enhancement. Given a low light image, in the first stage, we employ the tone mapping method proposed in [1] to pre-enhance the image. In the second stage, we use a refinement network for further improvement of image quality.

data. There are also many approaches have been proposed for denoising of low light images. Remez et al.[25] proposed a method that utilizes deep convolutional neural networks for Poisson denoising for low light images. Chatterjee [3] et al. used a locally linear embedding framework where a linear embedding is learned for denoising. It's noted that although previous approaches have demonstrated satisfactory performance for any of the above problems, it would be a difficult challenge when attempting to tackle them at the same time. We argue that simultaneously enhancing illumination as well as denoising is a non-trivial problem as they are usually formulated and solved in different paradigms.

To alleviate the above difficulties, in this paper, we decompose the low light into two sub-problems, i.e., the pre-enhancement and post-refinement, and propose a two-stage method to more accurately enhance low light images. To be specific, in the first stage, we enhance the illumination map decomposed from a low light image based on the Retinex theory. We employ a tone mapping based method [1] for the purpose. In the second stage, we design a refinement network to further improve the image quality from the pre-enhanced image obtained in the first stage. We design a comprehensive loss function that combines the loss of image content, perceptual quality, total variation, and adversarial loss. This stage contributes to the improvement of image quality, especially for noise suppression. Our two-stage strategy demonstrates satisfactory performance even for dark image inputs, an example is given in Fig.1 where the results are even better than the ground truth.

To summarize, the main contribution of this paper is the proposal of a simple two-stage unsupervised approach that performs pre-enhancement and post-refinement for low light image enhancement. It outperforms state-of-the-art methods, including both supervised learning based methods and unsupervised learning based methods on four benchmark datasets. Furthermore, we show two applications of our method in which we demonstrate that it can archive much accurate feature points matching and can be further seamlessly applied to SLAM in low light conditions.

II. Related Works

A. Traditional Methods

Traditional approaches can basically be separated into two categories: histogram equalization based methods and Retinex theory based methods. Among them, histogram equalization [6], [37] are the most simply and widely used methods. There are also many Retinex based approaches. Guo et al. proposed a method called LIME [9] which first initializes the illumination map with the maximum value in its RGB channels, then imposes a structure prior on the illumination map. [16] proposed a robust Retinex model that formulates low light image enhancement as an optimization problem. They additionally applied the l_1 norm on the illumination map to constrain the piece-wise smoothness of the illumination. However, these traditional approaches tend to cause color distortion and amply noise in enhanced images.

B. Supervised Based Methods

[29] proposed to learn a deep convolutional network that directly formulates the low light image enhancement as a machine learning problem. The network is learned by penalizing the error between low light images and their corresponding normal light images. [26] proposed a two-stream framework which consists of a content stream network and an edge stream network. [5] proposed to use a neural network to decompose a low light image into two components, i.e. an illumination map and a reflectance map based on the Retinex theory, then the enhancement is applied on the two components with ground truth illumination and reflectance map. A similar idea is also adopted in [34], where a more accurate network is introduced. It's noted that supervised learning based methods have brought significant progress on the task, however, the need of image pairs of low light and normal light images for learning makes them hard to be applied to real-world scenarios.

C. Unsupervised Based Methods

Unsupervised based methods attempt to enhance low light images without pairs of low light and normal light images. To this end, [17] proposed a deep auto-encoder based approach that learns to enhance from low light images in an unsupervised fashion where the low light images are synthesized with different dark conditions. Previous methods have also attempted to utilize generative adversarial network (GAN). [14] proposed EnlightenGAN which can be trained in an end-to-end fashion, it achieved competitive performance compared with supervised learning based methods. [33] further proposed decoupled networks where illumination enhancement and noise reduction are handled with contrast enhancement and image denoising network, respectively. Besides, Zhang et al.[35] assumed that the maximum channel of the reflectance should conform to the maximum channel of the low light image and has the maximum entropy. Based on the assumption, they introduced a maximum entropy based Retinex model which can be trained with low light images only. However, the method didn't demonstrate competitive performance against others such as EnlightenGAN.

III. Methodology

As discussed above, it's difficult to get satisfactory performance by directly formulating the low light image enhancement as a learning problem considering the difficulty of simultaneous illumination enhancement and denoising. Therefore, we propose a two-stage framework that performs pre-enhancement and post-refinement to gain better performance. The proposed framework is shown in Fig. 2. Given a low light image, we first enhance an illumination map decomposed from the low light input. Then the pre-enhanced image is inputted to a refinement network to further suppress noise and improve the overall quality. The details of our two-stage method are shown below.

A. *Pre-enhancement*

According to Retinex theory, an image can be decomposed into an illumination map and a reflectance map, i.e.,

$$X = I \circ R, \tag{1}$$

where X is an RGB image, I and R are illumination and reflectance map, respectively. In the first stage, we employ the adaptive tone mapping [1] to enhance the illumination map. It's represented as:

$$Y' = \frac{L_g}{L_w} \circ X \tag{2}$$

where Y' denotes the pre-enhanced image from X, L_w is the gray scale of X; L_g is the global adaptation output, it is calculated by:

$$L_g = \frac{\log(L_w/\overline{L}_w + 1)}{\log(L_{wmax}/\overline{L}_w + 1)}, \tag{3}$$

where L_{wmax} denotes the maximum of L_w. $\overline{L}_w$ is the log-average luminance which can be formulated as:

$$\overline{L}_w = \exp\left(\frac{1}{m*n}\sum(\log(\sigma + L_w))\right) \tag{4}$$

where m, n denotes the width and height of image, σ is a small constant number.

Note that the pre-enhancement can yield competitive performance compared with many deep learning based approaches in terms of illumination enhancement. However, on the other hand, it will largely amplify noise, as seen in the second row of Fig. 3. To cope with this problem, we employ a network for further refinement to improve image quality.

B. *Post-refinement*

The refinement network is an encoder-decoder network which is built on U-net [27]. The encoder consists of four convolutional layers, four downsampling layers. The downsampling layer consists of two convolutional layers followed by a max pooling layer. The encoder extracts features at multiple scales: $1/4$, $1/8$, $1/16$, and $1/32$. The decoder employs four upsampling layers to gradually up-scale the final features from the encoder and yields the final output with a convolutional layer. For upsampling, we employ the upsampling strategy used in [13], [11]. The details of the refinement net are given in Table I, where conv1 to conv5 are convolutional layers, down1 to down4 are downsampling layers, up1 to up4 are upsampling layers, respectively; Layers of fusion1 to fusion4 are used to concatenate and fuse the features of encoder layers and decoder layers at multi-scales. It consists of two convolutional layers.

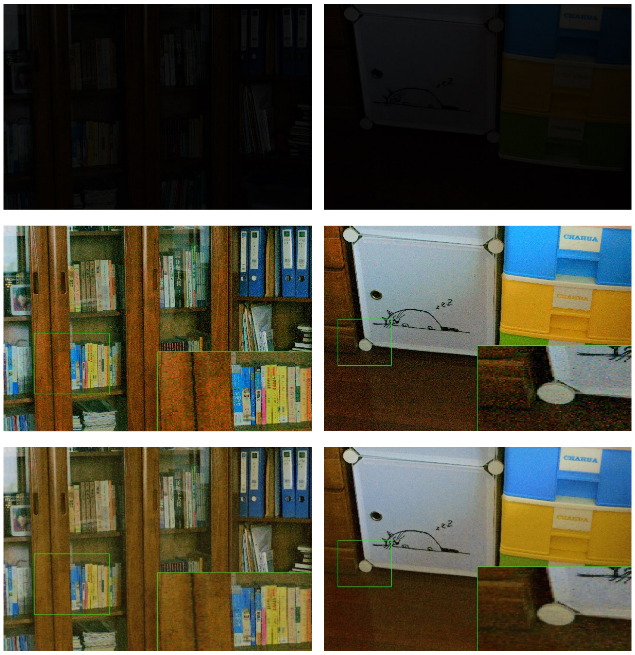


Fig. 3. The enhanced results from dark images. The first row denotes the original low light images, the second row shows the results of the pre-enhancement. The third row denotes the results of the two-stage method. It's clear that the post-refinement can effectively suppress noise.

TABLE I
Input/output, sizes of output features, and input/output channels of each layer for the refinement network on the training set of Unpaired Enhancement Dataset.

Layer	Input/Output	Output Size	Input/C	Output/C
conv1	Y'/x_1	128×128	3	32
conv2	x_1/x_2	128×128	32	32
down1	x_2/x_3	64×64	32	32
down2	x_3/x_4	32×32	32	64
down3	x_4/x_5	16×16	64	128
down4	x_5/x_6	8×8	128	256
conv3	x_6/x_7	8×8	256	512
conv4	x_7/x_8	8×8	512	512
up1	x_8/x_9	16×16	512	256
fusion1	$x_9, x_5/x_{10}$	16×16	384	256
up2	x_{10}/x_{11}	32×32	256	128
fusion2	$x_{11}, x_4/x_{12}$	32×32	192	128
up3	x_{12}/x_{13}	64×64	128	64
fusion3	$x_{13}, x_3/x_{14}$	64×64	96	64
up4	x_{14}/x_{15}	128×128	64	32
fusion4	$x_{15}, x_2/x_{16}$	128×128	64	32
conv5	x_{16}/Y	128×128	32	3

As the difficulty to obtain the paired images of low light and normal light in real-world applications, we design a comprehensive loss function that can be used to train the network in an unsupervised fashion. The loss function consists of four loss terms. The first term is a reconstruction loss that minimizes the pixel-wise loss of image. It ensures the consistency of image contents between the refined image and the pre-enhanced image. It is represented as:

$$l_{\mathrm{rec}} = \|Y - Y'\|_1, \tag{5}$$

where Y' denotes a pre-enhanced image, Y is a refined image from Y', it is calculated by the refinement network N, i.e. $Y = N(Y')$. In addition, we employ a perceptual loss to constrain the loss in feature space of VGG [30], it is represented as:

$$l_{\mathrm{per}} = \|\phi(Y) - \phi(Y')\|_2, \tag{6}$$

where ϕ denotes VGG network, $\phi(Y')$ is the feature maps extracted from Y'. The reconstruction loss and perceptual loss work in a complementary fashion to avoid color distortion and loss of image contents.

To suppress noise, we additionally apply total variation to the refined image,

$$l_{\mathrm{tv}} = \|\nabla Y\|_1, \tag{7}$$

l_{tv} contributes to the reduction of noise, however, it will also lead to the blurred effect on image structure. Therefore, we use an adversarial loss to encourage the refined image to be as close as the clear normal light image. Following [14], we use the relativistic discriminator structure [15] as the discriminative network which is fully convolutional and can handle the input with any size. Then the adversarial loss is given by:

$$l_{\mathrm{adv}} = ((D(Y) - D(\hat{Y})) - 1)^2 + (D(\hat{Y}) - D(Y))^2, \tag{8}$$

where D is the discriminator, $\hat{Y}$ denotes normal light images. As a result, the final loss function for training the refinement network is:

$$L = l_{\mathrm{rec}} + \lambda l_{\mathrm{per}} + \mu l_{\mathrm{tv}} + \beta l_{\mathrm{adv}}, \tag{9}$$

where λ, μ and β are weighting coefficients.

IV. EXPERIMENTS

A. Datasets

a) Unpaired Enhancement Dataset: The unpaired enhancement dataset [14] is collected from several public datasets. The training set is composed of 914 low light images and 1016 normal light images. The test set is composed of 148 pairs of low light and normal light images. All the images have a resolution of 600 × 400. We compare our method with the benchmark method [14], i.e. EnlightenGAN on this dataset.

b) Benchmark Evaluation Datasets: For a fair comparison with previous methods, we report more quantitative results on real-world benchmark datasets. We evaluate our method on MEF[19], LIME[9], NPE[31]. The three datasets are frequently used in previous studies for evaluation, in which MEF, LIME, and NPE have 17, 8, and 10 images, respectively.

TABLE II
QUANTITATIVE COMPARISONS ON THE UNPAIRED DATASET.

	PSNR ↑	SSIM↑	NIQE ↓
Input	10.370	0.275	5.299
EnlightenGAN [5]	17.314	0.711	4.591
Pre-enhancement [1]	17.337	0.698	7.012
Post-refinement	**18.064**	**0.720**	**4.474**

TABLE III
QUANTITATIVE COMPARISONS OF DIFFERENT METHODS ON THE BENCHMARK DATASETS.

	MEF	LIME	NPE
Input	4.265	4.438	4.319
RetinexNet [5]	4.149	4.420	4.485
LIME [9]	3.720	4.155	4.268
SRIE [8]	3.475	3.788	3.986
NPE [31]	3.524	3.905	3.953
GLAD [32]	3.344	4.128	3.970
EnlightenGAN [14]	3.232	3.719	4.113
KinD [34]	3.343	3.724	3.883
Ours	**3.027**	**3.599**	**3.014**

B. Implementation Details

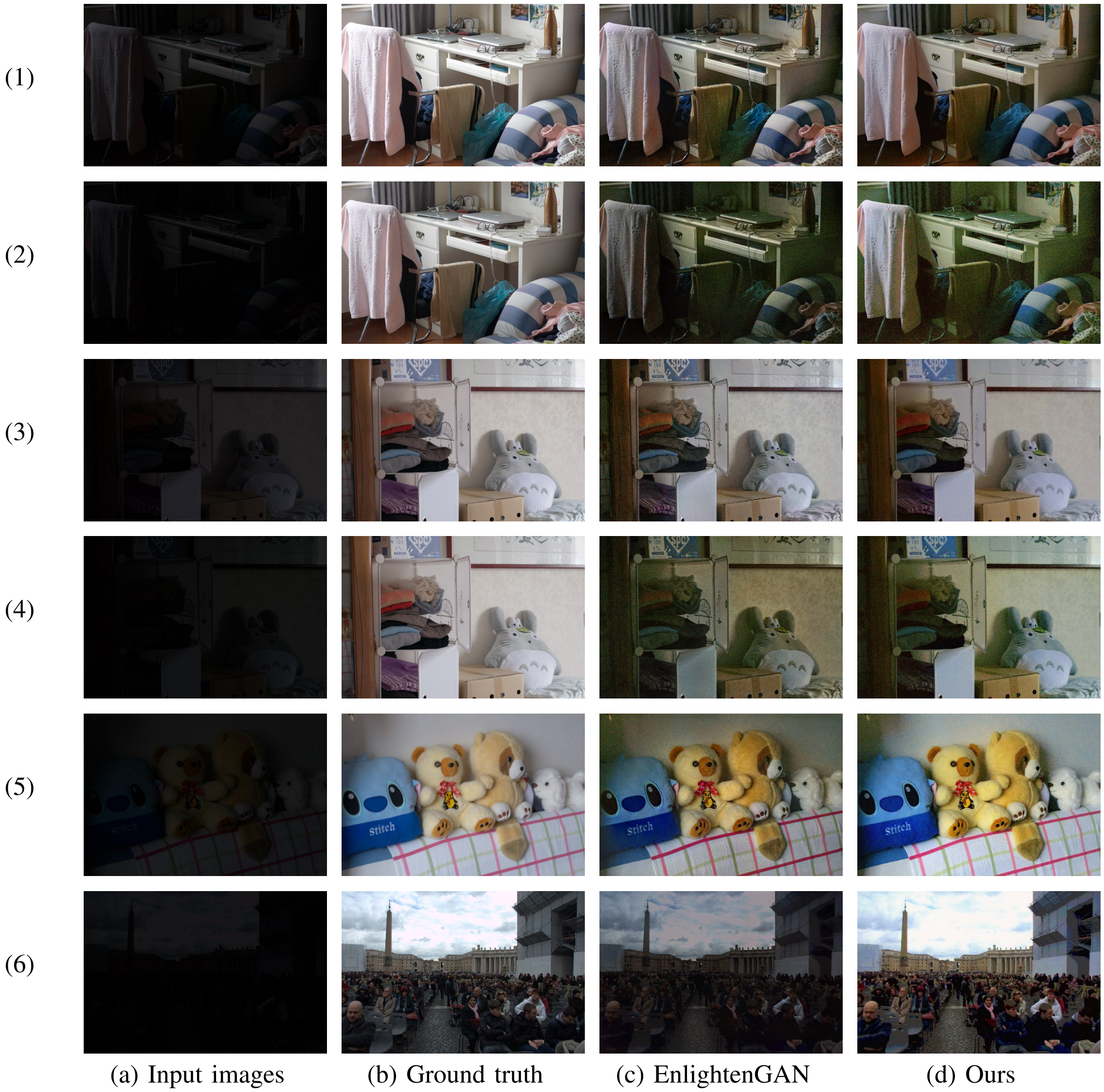
For learning the refinement network, we employ the training set from the unpaired enhancement dataset. During the training phase, we randomly crop 128×128 patches from the original 640×400 resolution images pixels.

We use Adam optimizer with a learning rate of 0.0001. We set $\beta_1 = 0.9$, $\beta_2 = 0.999$, and use weight decay of 0.0001. The weights λ of l_{per}, μ of l_{tv} and β of l_{adv} are set as $\lambda = 1$, $\mu = 0.01$ and $\beta = 1$, respectively, in all experiments throughout the paper. We train the refinement network for 1000 epochs. We conducted all the experiments using PyTorch [24] with batch size of 64.

C. Performance Comparison

We first show the quantitative comparison of our method against EnlightenGAN [14] on the unpaired enhancement dataset. Three metrics are adopted for quantitative comparison, which are PSNR, SSIM, and NIQE. For PSNR and SSIM, a higher value indicates a better quality, while for NIQE, the lower is better. As seen in Table II, the pre-enhancement yields a little better PSNR than EnlightenGAN, but the results are a little worse on SSIM, moreover, it is observed a 33.3% error increase of NIQE. On the other hand, our two-stage method achieves the best performance for all metrics, which indicates the superiority of the configuration of pre-enhancement and post-refinement. Fig. 4 shows the qualitative comparison against EnlightenGAN. It's observed that both EnlightenGAN and our method can achieve satisfactory performance if there are valuable clues that exist in the inputs as seen in Fig. 4 (1) and (3), however, it's difficult to get the same results if the inputs are extremely dark, as seen in Fig. 4 (2) and (4). Nevertheless, our method demonstrates better performance for dark images, as also seen in Fig. 4 (5) and (6).

For more comparisons against other methods, we provide the results of quantitative comparisons on the MEF, LIME, and



(a) Input images (b) Ground truth (c) EnlightenGAN (d) Ours

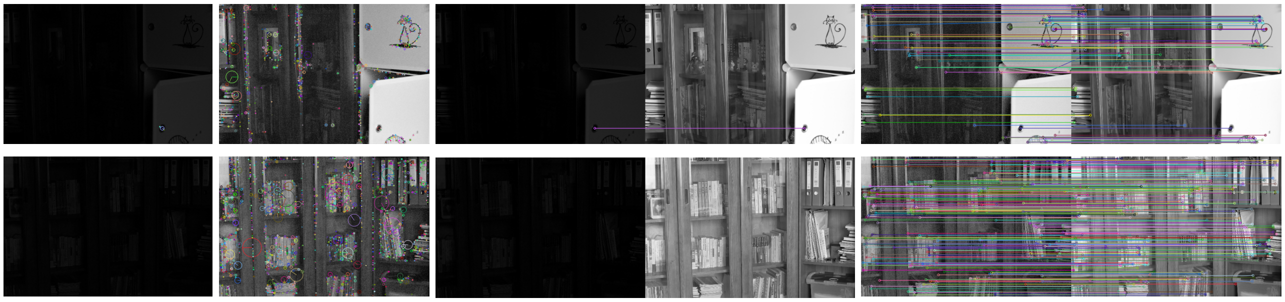
Fig. 4. Qualitative comparison between EnlightenGAN and our method on the unpaired enhancement dataset for different methods. From the left to the right; input low light images, ground truth images, results of EnlightenGAN, and results of our method. (1) - (2), (3) - (4) show the results for same scenes under different light conditions. (5) and (6) show results for other dark scenes. Our method outperforms EnlightenGAN for all eight images.

NPE datasets. It is noted that we do not train a new model for these three datasets. To evaluate the generability of our method, we use the trained model on the unpaired dataset and test it on the MEF, LIME, and NPE datasets. As there are no reference images are available for these datasets, we use the NIQE value as image quality evaluation in compliance with previous methods [14], [34]. We compare our method against RetinexNet [5], LIME [9],SRIE [8], NPE [31], GLAD [32], KinD [34], and EnlightenGAN [14]. The numerical results are shown in Table III, it is seen that our method shows a clear advantage against the others as it outperforms them for all datasets. Besides, we conduct an ablation study to compare the performance of different loss functions on the all mixed images of MEF, LIME and NPE. As a result, the NIQE is 3.187, 3.328, 3.203 and 3.586 for $l_{\rm rec}+l_{\rm per}+\mu l_{\rm tv}+l_{\rm adv}$, $l_{\rm per}+\mu l_{\rm tv}+l_{\rm adv}$, $l_{\rm per}+l_{\rm adv}$ and $l_{\rm per}$, respectively. Note that the weights λ, β for $l_{\rm per}$ and $l_{\rm adv}$ are set to 1, thus we omit them here. The experimental results show that the combination of full loss terms demonstrates the best enhancement performance.

D. Application: Low Light and Normal Light Image Matching

Image matching is one of the fundamental techniques in robot vision and it plays an indispensable role in many applications such as image retrieval, structure from motion, image based localization, etc. Unsurprisingly, the low light condition easily leads algorithms of feature points matching to malfunction. Zhou et al.[36] also discussed the necessity of image matching between a low light image and a normal light image.

We show that our method can be applied to low light and normal light image matching. To be specific, we conduct the image matching between a low light image and its corresponded normal light image on the test set of the unpaired dataset. We use SIFT to detect feature points and generate descriptors. Then, they are matched with the 2-nearest neighbor algorithm. To get more accurate matching, we use a



(a) Feature points from low light images (b) Feature points from normal light images (c) Low light - normal light image matching (d) Enhanced image - normal light image matching

Fig. 5. Qualitative results of feature points detection and matching. (a) shows results of detected points with SIFT from low light images, (b) shows results of detected points with SIFT from enhanced images of low light images, (c) is the results of image matching between low light images and normal light images, (d) is the results between enhanced images from low light images and normal light images.

TABLE IV
RESULTS FOR FEATURE POINTS DETECTION AND MATCHING WITH AND WITHOUT THE ENHANCEMENT OF LOW LIGHT IMAGES.

	Detected points	Matches	Match rate
Low light images	22195	17424	13.85%
EnlightenGAN	185930	38152	30.32%
Ours	172554	40676	32.33%

TABLE V
$SE3\_ATE\_RMSE(cm)$ ON $sfm\_lab\_room\_1$, $sfm\_lab\_room\_2$, $large\_loop\_1$ AND $plant\_scene\_1$.

	Original	EnlightenGAN	Ours
$sfm\_lab\_room\_1$	3.134	1.907	1.764
$sfm\_lab\_room\_2$	Fail	5.824	2.956
$large\_loop\_1$	Fail	10.401	4.552
$plant\_scene\_1$	Fail	3.356	1.428

small number for distance ratio. In our experiment, we set it to 0.3. To further eliminate mismatches, we apply the RANSAC algorithm [7] to remove outliers.

The quantitative results are given in Table IV. As a result, 22195 points are detected from the low light images (there are 125825 feature points detected from the normal light images). From that, we can only get 17424 matches, i.e. the match rate[1] is only 13.85%. On the other hand, after applying the enhancement with our method, the number of matched points is 40676 and the match rate is significantly improved from 13.85% to 32.33%. It's noted that EnlightenGAN detected more feature points than our method though, the match rate is lower than ours. It suggests that there are many noisy points detected by EnlightenGAN and the quality of enhanced images by EnlightenGAN are not as good as ours. A qualitative comparison is shown in Fig. 5 (c), as seen that there is almost no successful matching for the original low light images. This is because it's extremely difficult to detect feature points from low light images (Fig. 5 (a)). After applying low light image enhancement (Fig. 5 (b)), a large amount of feature points are detected and they can be correctly matched (Fig. 5 (d)).

E. Application: SLAM in Low Light Conditions

Vision based monocular SLAM tends to fail in low light environments. To evaluate the application of our method for SLAM, we test it on the ETH3D SLAM benchmark [28]. Specifically, we use the ORB-SLAM2 [22] to perform RGBD based monocular SLAM. We evaluate our method as well as EnlightenGAN $sfm\_lab\_room\_1$, $sfm\_lab\_room\_2$, $large\_loop\_1$ and $plant\_scene\_1$ taken from ETH3D SLAM benchmark [28]. They are captured in low light conditions but not completely dark. An example is given in Fig. 6, where the first row shows the original low light images taken from the above sequences, and the second row shows the images enhanced by our method.

Table V shows the quantitative comparisons. It's seen that ORB-SLAM2 only successfully performed on $sfm\_lab\_room\_1$[2] without the low light image enhancement. However it failed on $sfm\_lab\_room\_2$, $large\_loop\_1$ and $plant\_scene\_1$ which is consistent with the results given in the benchmark [28]. On the other hand, the SLAM can be improved significantly if we apply low light image enhancement. As seen that our method performs better than EnlightenGAN for all of these four sequences. It is slightly better on $sfm\_lab\_room\_1$ and outperforms EnlightenGAN on $sfm\_lab\_room\_2$, $large\_loop\_1$ and $plant\_scene\_1$ by a good margin (achieving 49.50%, 56.23% and 57.45% improvement of the accuracy). In our experiments, our method takes 95 ms to enhance a 739×458 resolution image on a computer with Intel(R) Xeon(R) CPU E5-2690 v3 and a GT1080Ti GPU card. Fig. 7 shows camera trajectories for different inputs. As there are several pieces of ground truth trajectories are missing for the sequences of $plant\_scene\_1$ and $large\_loop\_1$, we only show the correct ground truths. They are shown in green and the results from the original images, enhanced images by EnlightenGAN and our method are shown in red, orange and blue, respectively.

[1] the match rate is the rate of the number of final matches divided by points detected from normal light images.

[2] The $SE3\_ATE\_RMSE$ is 1.850 given in the official benchmark while the result is 3.134 performed by us. The reason is considered as the difference of setting of parameters for ORB-SLAM2.

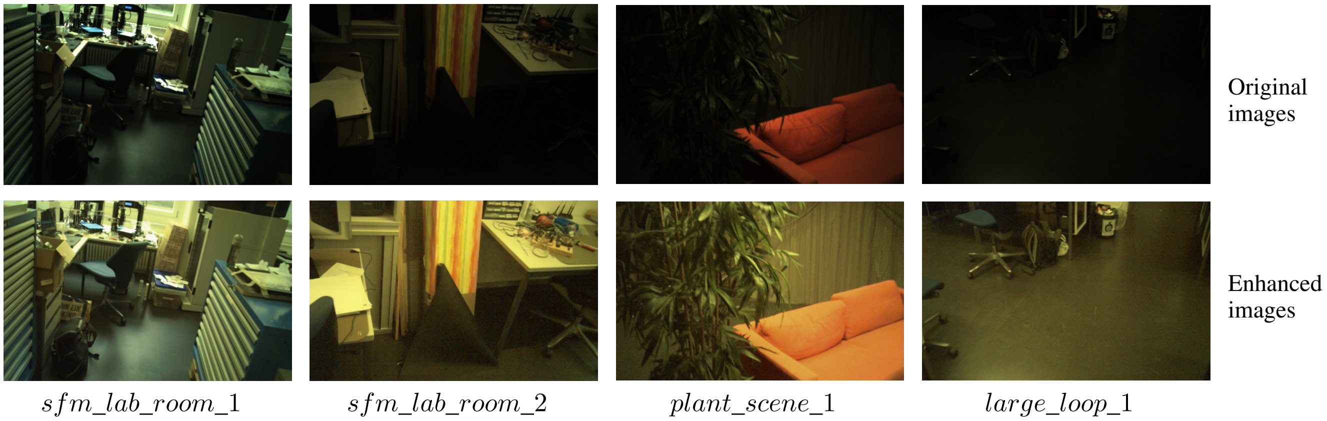


Fig. 6. Selected images from ETH3D SLAM benchmark dataset. The first row shows the original low light images, the second row shows the enhanced images with our method.

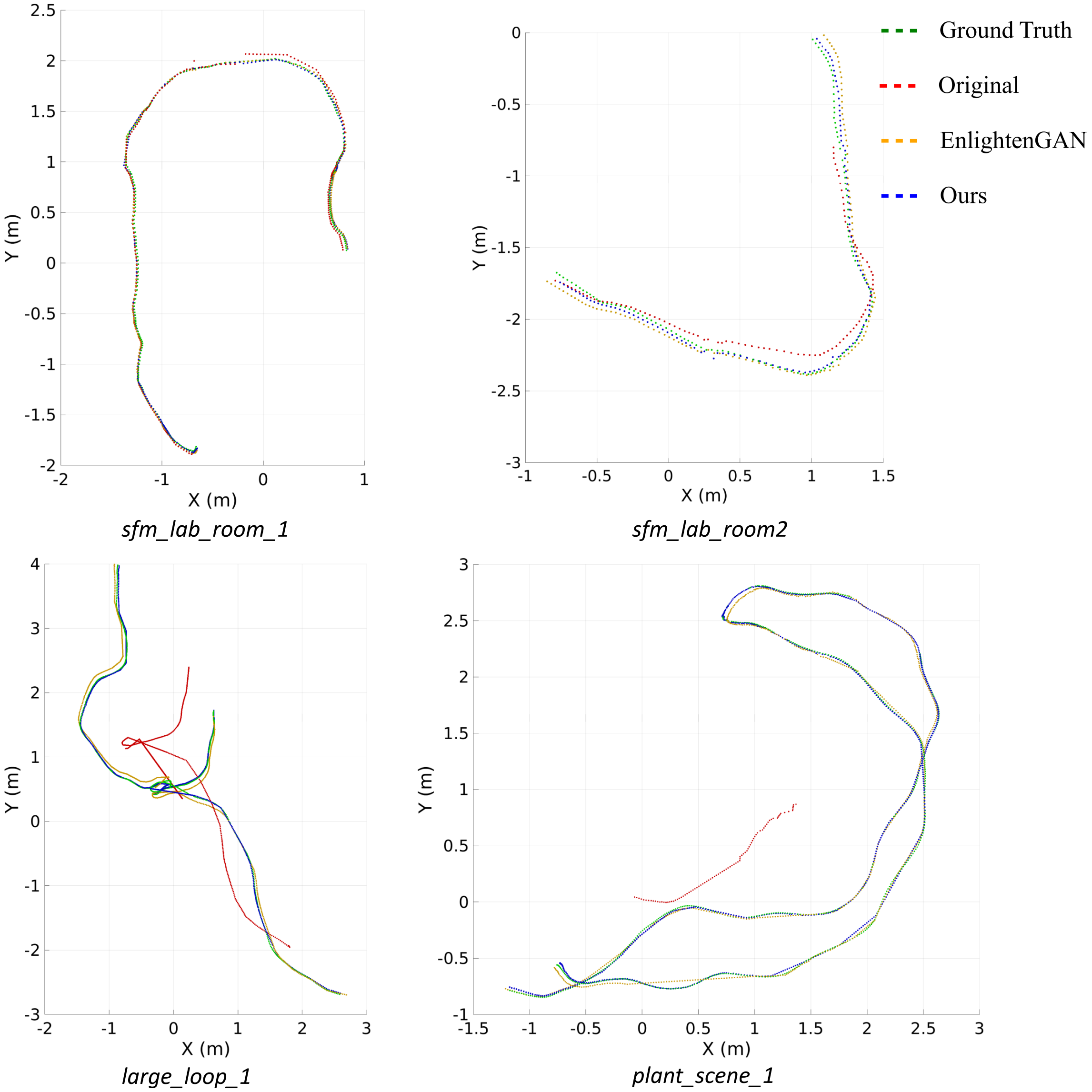


Fig. 7. Camera trajectories for the four sequences of ETH3D dataset, where the ground truths are shown in green and the results of our method are shown in blue.

V. Conclusion

In this paper, we revisited the problem of real-world low light image enhancement. We point out that there are mainly three challenges that hinder the deployment of most of previous learning based methods. The first is the need of low light and normal light image pairs for learning. To overcome this difficulty, we proposed an unsupervised method that can be implemented with unpaired images using adversarial training. Other challenges are the difficulty of handling very dark input images and the poor ability of denoising. To alleviate the difficulties, we take a two-stage strategy that first pre-enhances a low light image and further refines it with a refinement network. Experimental results show that our two-stage approach outperforms previous methods on four benchmark datasets. We argue that the proposed method can be used as an effective image pre-processing tool for low light image enhancement. In experiments, we demonstrate two useful applications of our method. The first is image matching and the second is SLAM. We show that both of them are vulnerable to low light conditions, nevertheless, they can be significantly improved with the image enhancement performed by our method. In the future, we will speed up our method with some compression techniques for deep neural networks and explore more applications for the perception of robots.

References

[1] H. Ahn, B. Keum, D. Kim, and H. S. Lee, “Adaptive local tone mapping based on retinex for high dynamic range images,” *2013 IEEE International Conference on Consumer Electronics (ICCE)*, pp. 153–156, 2013.

[2] G. Bresson, Z. Alsayed, L. Yu, and S. Glaser, “Simultaneous localization and mapping: A survey of current trends in autonomous driving,” *IEEE Transactions on Intelligent Vehicles*, vol. 2, no. 3, pp. 194–220, 2017.

[3] P. Chatterjee, N. Joshi, S. B. Kang, and Y. Matsushita, “Noise suppression in low-light images through joint denoising and demosaicing,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2011, pp. 321–328.

[4] C. Chen, Q. Chen, J. Xu, and V. Koltun, “Learning to see in the dark,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 3291–3300.

[5] W. Y. Chen Wei, Wenjing Wang and J. Liu, “Deep retinex decomposition for low-light enhancement,” in *British Machine Vision Conference*. British Machine Vision Association, 2018.

[6] D. Coltuc, P. Bolon, and J.-M. Chassery, “Exact histogram specification,” *IEEE Transactions on Image Processing*, vol. 15, pp. 1143–1152, 2006.

[7] M. A. Fischler and R. C. Bolles, “Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography,” *Communications of the ACM*, vol. 24, no. 6, pp. 381–395, 1981.

[8] X. Fu, D. Zeng, Y. Huang, X. Zhang, and X. Ding, “A weighted variational model for simultaneous reflectance and illumination estimation,” *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2782–2790, 2016.

[9] X. Guo, Y. Li, and H. Ling, “Lime: Low-light image enhancement via illumination map estimation,” *IEEE Transactions on Image Processing*, vol. 26, pp. 982–993, 2017.

[10] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[11] J. Hu, M. Ozay, Y. Zhang, and T. Okatani, “Revisiting single image depth estimation: Toward higher resolution maps with accurate object boundaries,” 2019.

[12] J. Hu, Y. Zhang, and T. Okatani, “Visualization of convolutional neural networks for monocular depth estimation,” in *IEEE International Conf. on Computer Vision (ICCV)*, 2019.

[13] L. Iro, R. Christian, B. Vasileios, T. Federico, and N. Nassir, “Deeper depth prediction with fully convolutional residual networks,” in *3DV*, 2016, pp. 239–248.

[14] Y. Jiang, X. Gong, D. Liu, Y. Cheng, C. Fang, X. Shen, J. Yang, P. Zhou, and Z. Wang, “Enlightengan: Deep light enhancement without paired supervision,” *arXiv preprint arXiv:1906.06972*, 2019.

[15] A. Jolicoeur-Martineau, “The relativistic discriminator: a key element missing from standard gan,” *arXiv preprint arXiv:1807.00734*, 2018.

[16] M. Li, J. Liu, W. Yang, X. Sun, and Z. Guo, “Structure-revealing low-light image enhancement via robust retinex model,” *IEEE Transactions on Image Processing*, vol. 27, pp. 2828–2841, 2018.

[17] K. G. Lore, A. Akintayo, and S. Sarkar, “Llnet: A deep autoencoder approach to natural low-light image enhancement,” *Pattern Recognition*, vol. 61, pp. 650–662, 2017.

[18] F. Ma and S. Karaman, “Sparse-to-dense: Depth prediction from sparse depth samples and a single image,” *ICRA*, 2018.

[19] K. Ma, K. Zeng, and Z. Wang, “Perceptual quality assessment for multi-exposure image fusion,” *IEEE Transactions on Image Processing*, vol. 24, pp. 3345–3356, 2015.

[20] A. Milan, T. Pham, K. Vijay, D. Morrison, A. W. Tow, L. Liu, J. Erskine, R. Grinover, A. Gurman, T. Hunn *et al.*, “Semantic segmentation from limited training data,” in *2018 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2018, pp. 1908–1915.

[21] A. Milioto, P. Lottes, and C. Stachniss, “Real-time semantic segmentation of crop and weed for precision agriculture robots leveraging background knowledge in cnns,” in *2018 IEEE international conference on robotics and automation (ICRA)*. IEEE, 2018, pp. 2229–2235.

[22] R. Mur-Artal and J. D. Tardós, “Orb-slam2: An open-source slam system for monocular, stereo, and rgb-d cameras,” *IEEE Transactions on Robotics*, vol. 33, no. 5, pp. 1255–1262, 2017.

[23] V. Nekrasov, T. Dharmasiri, A. Spek, T. Drummond, C. Shen, and I. Reid, “Real-time joint semantic segmentation and depth estimation using asymmetric annotations,” in *2019 International Conference on Robotics and Automation (ICRA)*. IEEE, 2019, pp. 7101–7107.

[24] A. Paszke, S. Gross, and A. Lerer, “Automatic differentiation in pytorch,” 2017.

[25] T. Remez, O. Litany, R. Giryes, and A. M. Bronstein, “Deep convolutional denoising of low-light images,” *arXiv preprint arXiv:1701.01687*, 2017.

[26] W. Ren, S. Liu, L. Ma, Q. Xu, X. Xu, X. Cao, J. Du, and M.-H. Yang, “Low-light image enhancement via a deep hybrid network,” *IEEE Transactions on Image Processing*, vol. 28, no. 9, pp. 4364–4375, 2019.

[27] O. Ronneberger, P. Fischer, and T. Brox, “U-net: Convolutional networks for biomedical image segmentation,” in *MICCAI*, 2015.

[28] T. Schöps, T. Sattler, and M. Pollefeys, “Bad slam: Bundle adjusted direct rgb-d slam,” *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 134–144, 2019.

[29] L. Shen, Z. Yue, F. Feng, Q. Chen, S. Liu, and J. Ma, “Msr-net: Low-light image enhancement using deep convolutional network,” *ArXiv*, vol. abs/1711.02488, 2017.

[30] K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” *arXiv preprint arXiv:1409.1556*, 2014.

[31] S. Wang, J. Zheng, H. Hu, and B. Li, “Naturalness preserved enhancement algorithm for non-uniform illumination images,” *IEEE Transactions on Image Processing*, vol. 22, pp. 3538–3548, 2013.

[32] W. Wang, C. Wei, W. Yang, and J. Liu, “Gladnet: Low-light enhancement network with global awareness,” *2018 13th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2018)*, pp. 751–755, 2018.

[33] W. Xiong, D. Liu, X. Shen, C. Fang, and J. Luo, “Unsupervised real-world low-light image enhancement with decoupled networks,” *arXiv preprint arXiv:2005.02818*, 2020.

[34] Y. Zhang, J. Zhang, and X. Guo, “Kindling the darkness: A practical low-light image enhancer,” in *Proceedings of the 27th ACM International Conference on Multimedia*, 2019, pp. 1632–1640.

[35] Y. Zhang, X. Di, B. Zhang, and C. Wang, “Self-supervised image enhancement network: Training with low light images only,” *arXiv*, pp. arXiv–2002, 2020.

[36] H. Zhou, T. Sattler, and D. W. Jacobs, “Evaluating local features for day-night matching,” in *European Conference on Computer Vision*. Springer, 2016, pp. 724–736.

[37] T. Çelik and T. Tjahjadi, “Contextual and variational contrast enhancement,” *IEEE Transactions on Image Processing*, vol. 20, pp. 3431–3441, 2011.