Low-Light Image Enhancement with Normalizing Flow

Yufei Wang, [1] **Renjie Wan,** [1] **Wenhan Yang,** [1] **Haoliang Li,** [2] **Lap-Pui Chau,** [1] **Alex C. Kot,** [1]

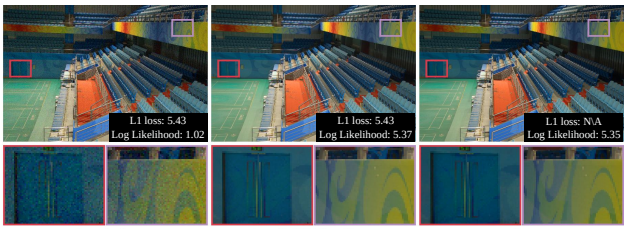
[1] Rapid-Rich Object Search Lab, Nanyang Technological University, Singapore
[2] Department of Electrical Engineering, City University of Hong Kong, China
{yufei,rjwan, wenhan.yang,elpchau,eackot}@ntu.edu.sg, haoliang.li@cityu.edu.hk

Abstract

To enhance low-light images to normally-exposed ones is highly ill-posed, namely that the mapping relationship between them is one-to-many. Previous works based on the pixel-wise reconstruction losses and deterministic processes fail to capture the complex conditional distribution of normally exposed images, which results in improper brightness, residual noise, and artifacts. In this paper, we investigate to model this one-to-many relationship via a proposed normalizing flow model. An invertible network that takes the low-light images/features as the condition and learns to map the distribution of normally exposed images into a Gaussian distribution. In this way, the conditional distribution of the normally exposed images can be well modeled, and the enhancement process, *i.e.*. the other inference direction of the invertible network, is equivalent to being constrained by a loss function that better describes the manifold structure of natural images during the training. The experimental results on the existing benchmark datasets show our method achieves better quantitative and qualitative results, obtaining better-exposed illumination, less noise and artifact, and richer colors.

1 Introduction

Low-light image enhancement aims to improve the visibility of low-light images and suppress captured noise and artifacts. Deep learning-based methods (Zhang et al. 2019; Zamir et al. 2020; Chen et al. 2018) achieve promising performance by utilizing the power of large collections of data. However, most of them mainly rely on the pixel-wise loss functions (*e.g.*, l_1 or l_2) in the network training that derive a deterministic mapping between the low-light and normally exposed images. This enhancement paradigm encounters two issues. First, this pixel-wise loss cannot provide effective regularization on the local structures in diverse contexts. As one low-light image may correspond to several reference images with different brightness (Zhang et al. 2021), this pixel-to-pixel deterministic mapping is easily trapped into the "regression to mean" problem and obtains the results that are the fusion of several desirable ones, which inevitably leads to improperly exposed regions and artifacts. Second, due to the simplified assumption of the pixel-wise losses about the image distribution, these losses might fail in describing the real visual distance between the reference image and enhanced images in the image manifold as shown in Fig. 1, which further undermines the performance. Though the GAN-based scheme can partly alleviate this issue, these approaches require careful tuning during training (Wolf et al. 2021) and might overfit certain visual features or the properties of the training data.



(a) noisy image (b) image with slight brightness variance (c) reference image

Figure 1: Illustration of the superiority of our normalizing flow model in measuring the visual distance compared to l_1 loss for low-light image enhancement. Although (b) is more visually similar to (c), *i.e.*, reference image, than (a), their l_1 losses are the same. Benefiting from better capturing the complex conditional distribution of normally exposed images, our model can better capture the error distribution and therefore provide the measure results more consistent with human vision.

Recently, researchers have shown the effectiveness of normalizing flow in the field of computational photography. (Wolf et al. 2021; Lugmayr et al. 2020; Xiao et al. 2020) The normalizing flow is capable to learn a more complicated conditional distribution than the classical pixel-wise loss, which can well solve the above-mentioned two issues. Beyond previous CNN-based models that learn a deterministic mapping from the low-light image to an image with specific brightness, the normalizing flow learns to map the multi-modal image manifold into a latent distribution. Then, the loss enforced on the latent space equivalently constructs an effective constraint on the enhanced image manifold. It leads to better characterization of the structural details in various contexts and better measurement of the visual distance in terms of high-quality well-exposed images, which helps effectively adjust the illumination and suppress the image

artifacts. However, since the classical normalizing flow is biased towards learning image graphical properties such as local pixel correlations (Kirichenko, Izmailov, and Wilson 2020), it may fail to model some global image properties like the color saturation, which can undermine the performance when applying these methods for the low-light image enhancement problem.

To address the above issues, in this paper, we propose **LLFlow**, a flow-based low-light image enhancement method to accurately learn the local pixel correlations and the global image properties by modeling the distributions over the normally exposed images. As shown in Fig. 2, to merge the global image information into the latent space, instead of using standard Gaussian distribution as the prior of latent features, we propose to use the illumination-invariant color map as the mean value of the prior distribution. More specifically, the encoder is designed to learn a one-to-one mapping to extract the color map that can be regarded as the intrinsic attributes of the scene that do not change with illumination. Simultaneously, another component of our framework, the invertible network, is designed to learn a one-to-many mapping from a low-light image to distribution of normally exposed images. As such, we expect to achieve better low-light image enhancement performance through our proposed framework.

In summary, contributions can be concluded as follows.

- We propose a conditional normalizing flow to model the conditional distribution of normally exposed images. It equivalently enforces an effective constraint on the enhanced image manifold. Via better characterization of the structural details and better measurement of the visual distance, it better adjusts illumination as well as suppresses noise and artifacts.
- We further introduce a novel module to extract the illumination invariant color map inspired by the Retinex theory as the prior for the low-light image enhancement task, which enriches the saturation and reduces the color distortion.
- We conduct extensively experiments on the popular benchmark datasets to show the effectiveness of our proposed framework. The ablation study and related analysis show the rationality of each module in our method.

2 Related works

2.1 Low-light image enhancement

As an active research topic in the past several years, a large number of low-light image enhancement methods have been proposed. Early methods mainly utilize the Retinex theory to correct the image illumination and suppress the artifacts. Recently, with the emergence of deep learning schemes, more tasks have benefited from the deep learning model. For example, LLNet (Lore et al. 2017) uses a deep auto-encoder to adaptively enlighten the image. Multi-scale features are adopted (Shen et al. 2017; Tao et al. 2017; Lv et al. 2018; Ren et al. 2019) to obtain better visual quality. The (Shen et al. 2017) illustrates the close relationship between Retinex and CNN with Gaussian convolution kernels, two separated deep networks are used for decomposition in (Wei et al. 2018), and (Wang et al. 2019b) propose a progressive Retinex framework that the illumination and reflection maps are trained in a mutually reinforced manner. In addition, different losses are used to guide the training, e.g., MSE (Lore et al. 2017; Cai, Gu, and Zhang 2018), l_1 loss(Cai, Gu, and Zhang 2018), structural similarity (SSIM) (Cai, Gu, and Zhang 2018), smoothness loss (Wang et al. 2019a; Zhang et al. 2019) and color loss (Wang et al. 2019a; Guo et al. 2020; Shen et al. 2017). Meanwhile, (Cai, Gu, and Zhang 2018) demonstrates that training the same network with different reconstruction losses will have different performances which demonstrates the significance of conditional distribution design. Introducing carefully designed color loss can be also regarded as refining the conditional distribution, *i.e.*, give color distortion pictures a greater penalty coefficient. Different from previous works that carefully design the reconstruction loss for end-to-end training, in this paper, we propose to utilize a normalizing flow to build the complex posterior distribution which has proven to be more effective and can generate images with higher quality, less noise, and artifact.

2.2 Normalizing flow

A normalizing flow is a transformation of a simple probability distribution (*e.g.*, a standard normal) into a more complex distribution by a sequence of invertible and differentiable mappings (Kobyzev, Prince, and Brubaker 2020). Meanwhile, the probability density function (PDF) value of a sample can be exactly obtained by transforming it back to the simple distribution. To make the network invertible and computation tractable, the layers of the network need to be carefully designed so that the inversion and the determinant of Jacbian matrix can be easily obtained which limits the capacity of the generative model. To this end, many powerful transformations have been proposed to enhance expressiveness capacity of the model. For example, affine coupling layers (Dinh, Krueger, and Bengio 2014), split and concatenation (Dinh, Krueger, and Bengio 2014; Dinh, Sohl-Dickstein, and Bengio 2016; Kingma and Dhariwal 2018), Permutation (Dinh, Krueger, and Bengio 2014; Dinh, Sohl-Dickstein, and Bengio 2016; Kingma and Dhariwal 2018), and 1×1 convolution (Kingma and Dhariwal 2018). Recently, conditional normalizing flows are investigated to improve the expressiveness of the model. (Trippe and Turner 2018) propose to use different normalizing flows for each condition. Recently, conditional affine coupling layer (Ardizzone et al. 2019; Winkler et al. 2019; Lugmayr et al. 2020) is used to build a stronger connection with the conditional feature and improve the efficiency of memory and computational resource. Benefiting from the development of normalizing flow, the scope of application has been greatly expanded. For instance, (Liu et al. 2019) generates faces with specific attributes, (Pumarola et al. 2020; Yang et al. 2019) use conditional flow to generate point clouds. In the super-resolution tasks, (Lugmayr et al. 2020; Winkler et al. 2019; Wolf et al. 2021) generate the distribution of high-resolution images based on one low-resolution input based on the conditional normalizing flow. Besides, the conditional normalizing flow is also used in image denoising (Abdelhamed, Brubaker, and Brown 2019; Liu et al. 2021b) to generate extra data or restore the clean image. In addi-

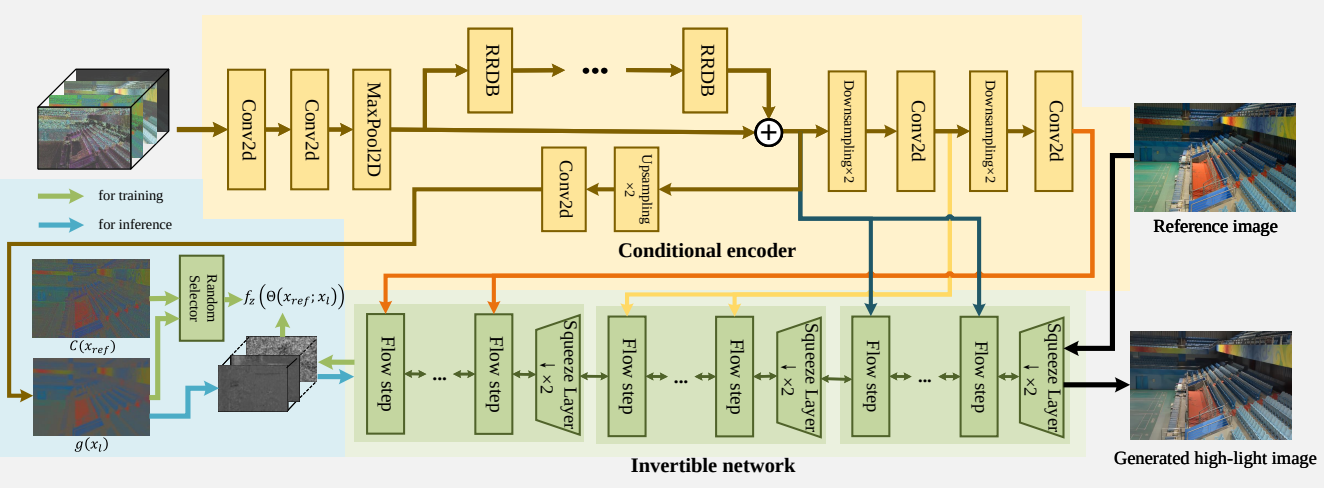


Figure 2: The architecture of our proposed LLFlow. Our model consists of a conditional encoder to extract the illumination-invariant color map and an invertible network that learns a distribution of normally exposed images conditioned on a low-light one. For training, we maximize the exact likelihood of a high-light image x_h by using change of variable theorem in Eq. (3) and a random selector is used to obtain the mean value of latent variable z which obey Gaussian distribution from the color map $C(x_h)$ of reference image or the extracted color map $g(x_l)$ from low-light image through the conditional encoder. For inference, we can randomly select z from $\mathcal{N}(g(x_l), \mathbf{1})$ to generate different normally exposed images from the learned conditional distribution $f_{flow}(x|x_l)$. (The color maps in the blue area are squeezed to the same size with latent feature z.)

tion, the inductive biases of normalizing flows are explored (Jaini et al. 2020; Kirichenko, Izmailov, and Wilson 2020). (Kirichenko, Izmailov, and Wilson 2020) reveals that the normalizing flow prefers to encode simple graphical structures which may be helpful to suppress the noise in the low-light image.

3 Methodology

In this section, we first introduce the limitations of previous pixel-wise reconstruction loss-based low-light enhancement methods. Then, the overall paradigm of our framework in Fig. 2 is introduced. Finally, two components of our proposed framework are illustrated separately.

3.1 Preliminary

The goal of low-light image enhancement is to generate a high-quality image with normal exposure x_h using a low-light image x_l. Paired samples (x_l, x_{ref}) are usually collected to train a model Θ by minimizing the l_1 reconstruction loss as follows:

$$\arg\min_{\Theta} \mathbb{E}\left[l_1(\Theta(x_l), x_{ref})\right] = \arg\max_{\Theta} \mathbb{E}\left[\log f(\Theta(x_l)|x_{ref})\right], \tag{1}$$

where $\Theta(x_l)$ is the normal-light image generated by the model and f is the probability density function conditioned on the reference image x_{ref} defined as follows:

$$f(x|x_{ref}) = \frac{1}{2b} \exp\left(-\frac{|x - x_{ref}|}{b}\right), \tag{2}$$

where b is a constant related to the learning rate. However, such a training paradigm has a limitation that the pre-defined distribution (*e.g.*, the distribution in Eq. 2) of images is not strong enough to distinguish between the generated realistic normally exposed image and the images with noises or artifacts such as the example in Fig. 1.

3.2 Framework

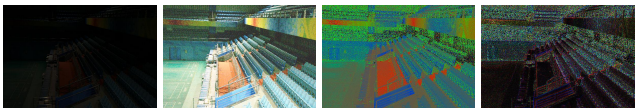
To this end, we propose to model the complicated distribution of normally exposed images using a normalizing flow so that the conditional PDF of a normally exposed image can be expressed as $f_{flow}(x|x_l)$. More specifically, a conditional normalizing flow Θ is used to take a low-light image itself and/or its features as input and maps a normally exposed image x to a latent code z which has the same dimension with x, *i.e.*, $z = \Theta(x; x_l)$. By using the change of variable theorem, we can obtain the relationship between $f_{flow}(x|x_l)$ and $f_z(z)$ as follows:

$$f_{flow}(x|x_l) = f_z\left(\Theta(x_{ref}; x_l)\right) \left|\det \frac{\partial \Theta}{\partial x_{ref}}(x_{ref}; x_l)\right|. \tag{3}$$

To make the model better characterize the properties of high-quality normally exposed images, we use the maximum likelihood estimation to estimate the parameter Θ. Specifically, we minimize the negative log-likelihood (NLL) instead of l_1 loss to train the model

$$\begin{aligned} L(x_l, x_{ref}) &= -\log f_{flow}(x_{ref}|x_l) \\ &= -\log f_z(\Theta(x_{ref}; x_l)) - \sum_{n=0}^{N-1} \log|\det \frac{\partial \theta^n}{\partial z^n}(z^n; g^n(x_l))|, \end{aligned} \tag{4}$$

where the invertible network Θ is divided into a sequence of N invertible layers $\{\theta^1, \theta^2, ..., \theta^N\}$ and $h^{i+1} = \theta^i(h^i; g^i(x_l))$ is the output of layer θ^i (i ranges from 0 to



(a) x_l (b) $h(x_l)$ (c) $C(x_l)$ (d) $N(x_l)$

Figure 3: The components of the input for the encoder g. The low-light image x_l, low-light image after histogram equalization $h(x_l)$, color map $C(x_l)$ and noise map $N(x_l)$ are concatenated to form the input with 12 channels.

$N-1$), $h^0 = x_{ref}$ and $z = h^N$. $g^n(x_l)$ is the latent feature from the encoder g that has the compatible shape with the layer θ^n. f_z is the PDF of the latent feature z.

In summary, our proposed framework includes two components: an encoder g which takes a low-light image x_l as input and output illumination invariant color map $g(x_l)$ (which can be regarded as reflectance map inspired by Retinex theory), and an invertible network that maps a normally exposed image to a latent code z. The details of the two components are introduced in the following subsections.

Encoder for illumination invariant color map: To generate robust and high quality illumination invariant color maps, the input images are first processed to extract useful features and the extracted features are then also concatenated as a part of the input of the encoder built by Residual-in-Residual Dense Blocks (RRDB) (Wang et al. 2018). The detailed architecture of the encoder g is in appendix due to limited space. The visualizations of each component are shown in Fig. 3 and the details are as follows:

1) Histogram equalized image $h(x_l)$: Histogram equalization is conducted to increase the global contrast of low-light images. The histogram equalized image can be regarded as a more illumination invariant one. By including the histogram equalized image as a part of the network's input, the network can better deal with the areas that are too dark or bright.

(a) $C(x_l)$ (b) $g(x_l)$ (c) $C(x_{ref})$ (d) x_{ref}

Figure 4: The color map directly extracted from the low-light image x_l, obtain from the encoder g, directly extracted from the reference image x_{ref}, and the reference image itself.

2) Color map $C(x)$: Inspired by Retinex theory, we propose to calculate the color map of an image x as follows:

$$C(x) = \frac{x}{\text{mean}_c(x)}, \tag{5}$$

where mean_c calculates the mean value of each pixel among RGB channels. The comparison between the color map from the low-light image, reference image, and the color map fine-tuned by the encoder g is shown in Fig. 4. As we can see, the color maps $C(x_l)$ and $C(x_{ref})$ are consistent to an extent under different illumination so they can be regarded as representations similar to the reflectance map, degraded with intensive noises in $C(x_l)$. We can also find that the encoder g can generate a high-quality color map that suppresses the strong noises to an extent and preserves the color information.

3) Noise map $N(x_l)$: To remove the noise in $C(x_l)$, a noise map $N(x_l)$ is estimated and fed into the encoder as an attention map. The noise map $N(x_l)$ is estimated as follows:

$$N(x) = \max(\text{abs}(\nabla_x C(x)), \text{abs}(\nabla_y C(x))), \tag{6}$$

where ∇_x, and ∇_y are the gradient maps in the directions of x and y, where $\max(x, y)$ is the operation that returns the maximum value between x and y at the pixel channel level.

Invertible network: Different from the encoder that learns a one-to-one mapping to extract illumination invariant color map which can be seen as the intrinsic invariant properties of the objects, the invertible network aims to learn a one-to-many relationship since the illumination may be diverse for the same scenario. Our invertible network is composed of three levels, and at each level, there are a squeeze layer and 12 flow steps. More details about the architecture can be found in the appendix.

According to our assumption that the normalizing flow aims to learn a conditional distribution of the normally exposed images conditioned on the low-light image/the illumination invariant color map, the normalizing flow should work well conditioned on both $g(x_l)$ and $C(x_{ref})$ since these two maps are expected to be similar. To this end, we train the whole framework (both the encoder and the invertible network) in the following manner:

$$\begin{aligned} L(x_l, x_{ref}) = &- \log f_z(\Theta(x_{ref}; x_l)) \\ &- \sum_{n=0}^{N-1} \log \left| \det \frac{\partial \theta^n}{\partial z^n}(z^n; g^n(x_l)) \right|, \end{aligned} \tag{7}$$

where f_z is the PDF of the latent feature z defined as follows

$$f_z(z) = \frac{1}{\sqrt{2\pi}} \exp\left(\frac{-(x - r(C(x_{ref}), g(x_l)))^2}{2} \right) \tag{8}$$

and $r(a, b)$ is a random selection function that is defined as follows:

$$r(a, b) = \begin{cases} a & \alpha \leq p \\ b & \alpha > p \end{cases}, \alpha \sim U(0, 1), \tag{9}$$

in which p is a hyper-parameter and we set p to be 0.2 for all experiments. As shown in Fig. 4, even without the help of pixel reconstruction loss, the encoder g can learn a similar color map with the reference image.

To generate a normally exposed image using a low-light image, the low-light image is first passed through the encoder to extract the color map $g(x_l)$ and then the latent features of the encoder are used as the condition for the invertible network. For the sampling strategy of z, one can randomly select a batch of z from the distribution $\mathcal{N}(g(x_L), \mathbf{1})$ to get

different outputs and then calculate the mean of generated normally-exposed images to achieve better performance. To speed up the inference, we directly select $g(x_l)$ as the input z and we empirically find that it can achieve a good enough result. So for all the experiments, we just use the mean value $g(x_l)$ as the latent feature z for the conditional normalizing flow if not specified.

4 Experiments

4.1 Experimental Settings

The patch size is set to 160×160 and the batch size is set to 16. We use Adam as the optimizer with a learning rate of 5×10^{-4} and without weight decay. For LOL dataset, we train the model for 3×10^4 iterations and the learning rate is decreased with a factor of 0.5 at 1.5×10^4, 2.25×10^4, 2.7×10^4, 2.85×10^4 iterations. For VE-LOL dataset, we train the model for 4×10^4 iterations and the learning rate is decreased with a factor of 0.5 at 2×10^4, 3×10^4, 3.6×10^4, 3.8×10^4 iterations.

4.2 Evaluation on LOL

We first evaluate our method on the LOL datset (Wei et al. 2018) including 485 images for training and 15 images for testing. Three metrics are adopted for quantitative comparison including PSNR, SSIM (Wang et al. 2004), and LPIPS (Zhang et al. 2018). The numerical results among different methods are reported in Table 1. From Table 1, we can find that our method significantly outperforms all the other competitors. The higher PSNR values show that our method is capable of suppressing the artifacts and better recovering color information. The better SSIM values demonstrate that our method better preserves the structural information with high-frequency details. In terms of LPIPS, a metric designed for the human perception, our method also achieves the best performance, which indicates our method better align with the human perception. The qualitative results are shown in Fig. 5 Our method achieves more promising perceptual quality by better suppressing the artifacts and revealing image details.

Table 1: Quantitative comparison on the LOL dataset (Wei et al. 2018) in terms of PSNR, SSIM and LPIPS. ↑ (↓) denotes that, larger (smaller) values lead to better quality.

Method	PSNR ↑	SSIM ↑	LPIPS ↓
Zero-DCE (Guo et al. 2020)	14.86	0.54	0.33
LIME (Guo, Li, and Ling 2016)	16.76	0.56	0.35
EnlightenGAN (Jiang et al. 2021)	17.48	0.65	0.32
RetinexNet (Wei et al. 2018)	16.77	0.56	0.47
RUAS (Risheng et al. 2021)	18.23	0.72	0.35
DRBN (Yang et al. 2020)	20.13	0.83	0.16
(Lv, Li, and Lu 2021)	20.24	0.79	0.14
KinD (Zhang et al. 2019)	20.87	0.80	0.17
KinD++ (Zhang et al. 2021)	21.30	0.82	0.16
LLFlow (Ours)	**25.19**	**0.93**	**0.11**

4.3 Evaluation on VE-LOL

To better evaluate the performance and generality of our method, we further perform evaluation on VE-LOL (Liu et al. 2021a) dataset. It is a large-scale dataset including 2500 paired images with more diversified scenes and contents, thus is valuable for the cross-dataset evaluation.

1) Cross-dataset evaluation: We first evaluate the generality of our method in a cross-dataset manner, *i.e.*, we train our method on the LOL dataset (Wei et al. 2018) and test the model on the testing set of VE-LOL dataset (Liu et al. 2021a). The quantitative results are reported in Table 2. From the results, our method significantly outperforms other methods in terms of all metrics. The qualitative comparisons of real-captured image are given in Fig. 7. The results generated by our methods are with less noise and better color saturation.

Table 2: Quantitative comparison on the VE-LOL dataset in terms of PSNR, SSIM and LPIPS. The models are trained on the training set of LOL. ↑ (↓) denotes that, larger (smaller) values lead to better quality.

Method	PSNR ↑	SSIM ↑	LPIPS ↓
RetinexNet (Wei et al. 2018)	14.68	0.5252	0.6423
BIMEF (Ying, Li, and Gao 2017)	15.95	0.6386	0.4573
DeepUPE (Wang et al. 2019a)	13.19	0.4902	0.4634
JED (Ren et al. 2018)	16.73	0.6817	0.3899
LIME (Guo, Li, and Ling 2016)	14.07	0.5274	0.4021
SICE (Cai, Gu, and Zhang 2018)	18.06	0.7094	0.5078
LLNet (Lore et al. 2017)	17.57	0.7388	0.4021
SRIE (Fu et al. 2016)	13.66	0.5509	0.4577
KinD (Zhang et al. 2019)	18.42	0.7658	0.2879
KinD++ (Zhang et al. 2021)	17.63	0.7994	0.2257
Zero-DCE (Guo et al. 2020)	21.12	0.7705	0.2480
EnlightenGAN (Jiang et al. 2021)	20.43	0.7921	0.2416
LLFlow (Ours)	**23.85**	**0.8986**	**0.1456**

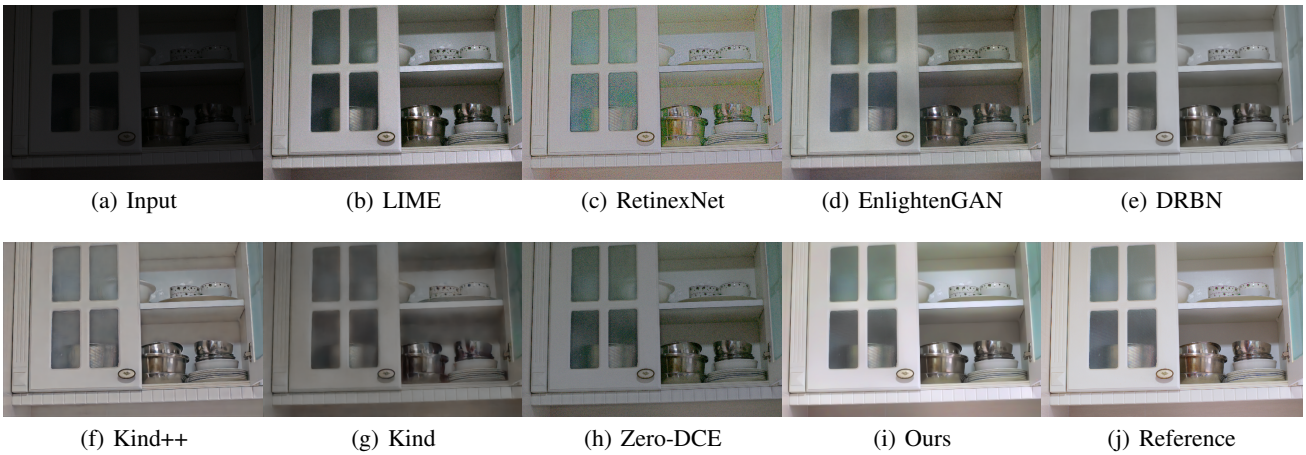
2) Intra-dataset evaluation: To further evaluate the performance of our proposed model, we compare our method with SOTA methods in an intra-dataset setting, *i.e.*, we retrain all the methods using the training set of VE-LOL dataset and report the performance on its corresponding test set. The quantitative results are reported in Table 3. We can find that our method has the best performance and outperforms others by a large margin. Meanwhile, with the help of more diverse data, all the metrics of our method are improved comparing with the model trained on LOL.

Table 3: Quantitative comparison on the VE-LOL dataset in terms of PSNR, SSIM, and LPIPS. The models are re-trained on the training set of VE-LOL dataset. ↑ (↓) denotes that, larger (smaller) values lead to better quality.

Method	PSNR ↑	SSIM ↑	LPIPS ↓
Zero-DCE (Guo et al. 2020)	20.54	0.7786	0.3312
KinD (Zhang et al. 2019)	22.15	0.8535	0.2576
LLFlow (Ours)	**26.02**	**0.9266**	**0.0996**

4.4 Ablation study

1) **The losses estimated by our method and l_1:** To verify our motivation that conditional normalizing flow can model a more complicated error distribution comparing with pixel-wise reconstruction loss, we further compare the losses obtained by our method and l_1. As shown in Table 4, the image with intensive noises and that with slightly different brightness have the same likelihood values under the measurement



(a) Input (b) LIME (c) RetinexNet (d) EnlightenGAN (e) DRBN

(f) Kind++ (g) Kind (h) Zero-DCE (i) Ours (j) Reference

Figure 5: Visual comparison with state-of-the-art low-light image enhancement methods on LOL dataset. The normally exposed image generated by our method has less noise and artifact, and better colorfulness.

(a) (b) (c)

Figure 6: (a) and (b) are the generated normally exposed images with different z (monotonically, each column) from a well-trained model. There are strong artifacts in (c) obtained from an early checkpoint of the model when it cannot well distinguish the artifacts and the variance of data. Zoom in to see details.

of l_1 loss, while the latter has much higher likelihood values under the measurement of our model than the former, which is better aligned with human perception.

Table 4: The differences under l_1 and negative log likelihood (NLL) estimated by our method for images with brightness variance and strong noise. The mean values among the test set of the LOL dataset are reported in the table.

Degradation	NLL estimated by our method	l_1
Reference	-6.09	N/A
Brightness reduced by 20[1]	-5.95	20
Brightness increased by 20	-6.15	20
Reference + random noise r[2]	4.84	20

[1] The range of pixel value is $0 - 255$.
[2] The random noise r has the same shape with the reference image and the mean value and mean absolute value of r are 0 and 20 respectively.

2) **The effect of different z:** A major advantage of our method over existing ones is that LLFlow can better encode the brightness variance into the latent space z. To verify the effectiveness of such strategy, we add a constant to the extracted $g(x_l)$ from -0.4 to 0.4 with a step of 0.2. The results in Fig. 6 demonstrate that the brightness of the image is monotonous with the value of z, which indicates that our model can encode the variance of the dataset, *i.e.*, the inevitable uncertainty when collecting the data pairs.

3) **The activation area of LLFlow:** To better understand how our model builds a more strong constraint, we visualize the gradient activation map of our method. For a normally exposed image x_{high} which can be a reference image or the output from a low-light enhancement network from its corresponding low-light image x_{low}, the gradient activation map G can be obtained as follows:

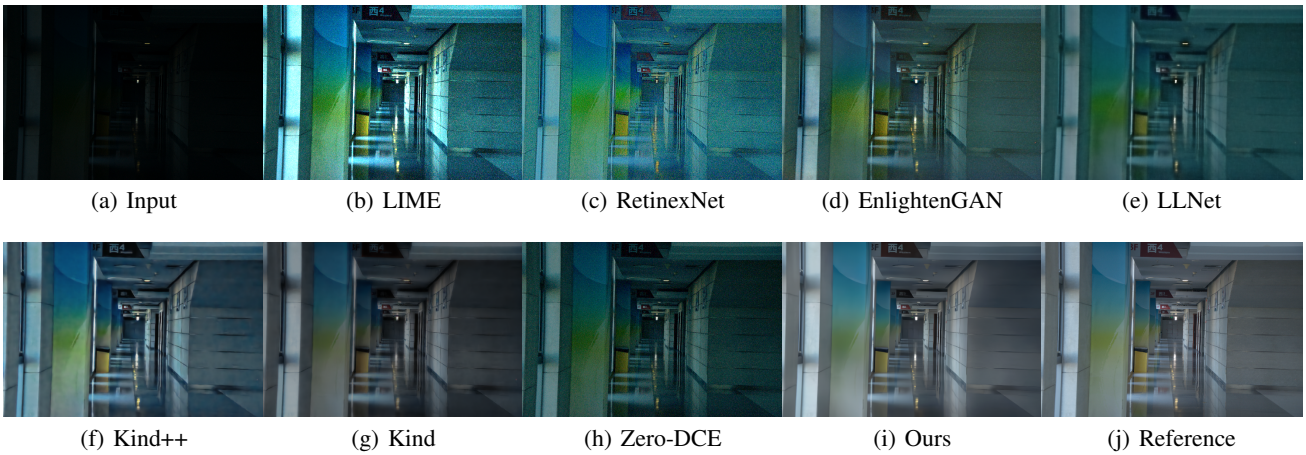
$$G = h(||\nabla_x L(x_l, x_{high})||_2) \tag{10}$$

where h is the histogram equalization operation to better visualize the results. From the results in Fig. 8, we can find that the area with artifacts has a higher gradient activation value. It demonstrates that even without the reference image, our model can distinguish the unrealistic areas according to the learned conditional distribution.

4) **The effectiveness of model components and training paradigm:** To investigate the effectiveness of our training paradigm and different components in our framework, We evaluate the performance of our conditional encoder individually and the performance of our whole framework via training them using l_1 loss.

For the evaluation of our whole framework under l_1 loss, we empirically find that training directly with it cannot converge. To this end, we first pretrain the framework for 1,000 iterations by minimizing the negative log likelihood $L(x_l, x_{ref})$. All the networks are trained with the same batch size, patch size, image prepossessing pipeline in the related experiments. We finetune other hyper-parameters, e.g., learning rate and weight decay, in a wide range to achieve the best performance.

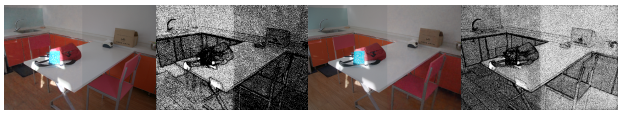
The results evaluated on LOL dataset (Wei et al. 2018) are reported in Table 5. The model trained by minimizing NLL loss has a huge improvement in all metrics comparing with the model trained by l_1 loss. A visual comparison between the results from l_1 loss trained model and NLL trained model



(a) Input (b) LIME (c) RetinexNet (d) EnlightenGAN (e) LLNet

(f) Kind++ (g) Kind (h) Zero-DCE (i) Ours (j) Reference

Figure 7: Visual comparison with state-of-the-art low-light image enhancement methods on the real-captured set of VE-LOL dataset.



(a) (b) (c) (d)

Figure 8: Gradient activation map from our model. **(a)**: The stitched picture that its left half is from not fully trained model and its right half is from the reference image. **(b)**: The gradient activation map from (a). **(c)**: The stitched picture that its right half is from not fully trained model and its left half is the reference image. **(d)**: The gradient activation map from (c).

are shown in Fig. 9. From the results, the model trained by l_1 loss produces more obvious artifacts. Both quantitative and qualitative results demonstrate the superiority of our flow-based method in modeling the distribution of images with normal brightness over a simplified pixel-wise loss.

(a) reference image (b) by NLL loss (c) by L1 loss

Figure 9: The effect of different training paradigm for the same network.

5) **The effect of different latent feature distributions:** To evaluate the effectiveness of our proposed illumination invariant color map and different hyper-parameters p, we evaluate them using the LOL dataset (Wei et al. 2018). The results in Table 6 show that our whole model with the newly designed color map achieves better PSNR values. The higher SSIM and LPIPS values show that the color map helps improve the color and brightness consistency.

Table 5: Quantitative comparison between training the model with l_1 and NLL loss on the LOL dataset. ↑ (↓) denotes that, larger (smaller) values lead to better quality.

Loss	PSNR ↑	SSIM ↑	LPIPS ↓
Only encoder (L1 loss)	21.90	0.8587	0.1672
LLFlow (L1 loss)	22.68	0.8391	0.2038
LLFlow (Ours)	**25.19**	**0.9252**	**0.1131**

Table 6: The effect of different latent feature distributions. ↑ (↓) denotes that, larger (smaller) values lead to better quality.

Latent Distribution	PSNR ↑	SSIM ↑	LPIPS ↓
LLFlow w/o color map	24.46	0.9235	0.1146
LLFlow w/ color map, $p = 0.5$	24.85	0.9232	0.1192
LLFlow w/ color map, $p = 0.2$	**25.19**	**0.9252**	**0.1131**

5 Conclusion

In this paper, we propose a novel framework for low-light image enhancement through a novel normalizing flow model. Compared with the existing techniques based on the pixel-wise reconstruction losses with deterministic processes, the proposed normalizing flow trained with negative log-likelihood (NLL) loss taking the low-light images/features as the condition naturally better characterizes the structural context and measures the visual distance in image manifold. With these merits, our proposed method naturally better captures the complex conditional distribution of normally exposed images and can achieve better low-light enhancement quality, *i.e.*, well-exposed illumination, suppressed noise and artifacts, as well as rich colors. The experimental results on the existing benchmark datasets show that our proposed framework can achieve better quantitative and qualitative results compared with state-of-the-art techniques.

References

Abdelhamed, A.; Brubaker, M. A.; and Brown, M. S. 2019. Noise flow: Noise modeling with conditional normalizing flows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 3165–3173.

Ardizzone, L.; Lüth, C.; Kruse, J.; Rother, C.; and Köthe, U. 2019. Guided image generation with conditional invertible neural networks. *arXiv preprint arXiv:1907.02392*.

Cai, J.; Gu, S.; and Zhang, L. 2018. Learning a deep single image contrast enhancer from multi-exposure images. *IEEE Transactions on Image Processing*, 27(4): 2049–2062.

Chen, C.; Chen, Q.; Xu, J.; and Koltun, V. 2018. Learning to see in the dark. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 3291–3300.

Dinh, L.; Krueger, D.; and Bengio, Y. 2014. Nice: Nonlinear independent components estimation. *arXiv preprint arXiv:1410.8516*.

Dinh, L.; Sohl-Dickstein, J.; and Bengio, S. 2016. Density estimation using real nvp. *arXiv preprint arXiv:1605.08803*.

Fu, X.; Zeng, D.; Huang, Y.; Zhang, X.-P.; and Ding, X. 2016. A weighted variational model for simultaneous reflectance and illumination estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2782–2790.

Guo, C.; Li, C.; Guo, J.; Loy, C. C.; Hou, J.; Kwong, S.; and Cong, R. 2020. Zero-reference deep curve estimation for low-light image enhancement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1780–1789.

Guo, X.; Li, Y.; and Ling, H. 2016. LIME: Low-light image enhancement via illumination map estimation. *IEEE Transactions on Image Processing*, 26(2): 982–993.

Jaini, P.; Kobyzev, I.; Yu, Y.; and Brubaker, M. 2020. Tails of lipschitz triangular flows. In *International Conference on Machine Learning*, 4673–4681. PMLR.

Jiang, Y.; Gong, X.; Liu, D.; Cheng, Y.; Fang, C.; Shen, X.; Yang, J.; Zhou, P.; and Wang, Z. 2021. Enlightengan: Deep light enhancement without paired supervision. *IEEE Transactions on Image Processing*, 30: 2340–2349.

Kingma, D. P.; and Dhariwal, P. 2018. Glow: Generative flow with invertible 1x1 convolutions. *arXiv preprint arXiv:1807.03039*.

Kirichenko, P.; Izmailov, P.; and Wilson, A. G. 2020. Why normalizing flows fail to detect out-of-distribution data. *arXiv preprint arXiv:2006.08545*.

Kobyzev, I.; Prince, S.; and Brubaker, M. 2020. Normalizing flows: An introduction and review of current methods. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Liu, J.; Xu, D.; Yang, W.; Fan, M.; and Huang, H. 2021a. Benchmarking Low-Light Image Enhancement and Beyond. *International Journal of Computer Vision*, 129(4): 1153–1184.

Liu, R.; Liu, Y.; Gong, X.; Wang, X.; and Li, H. 2019. Conditional adversarial generative flow for controllable image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7992–8001.

Liu, Y.; Qin, Z.; Anwar, S.; Ji, P.; Kim, D.; Caldwell, S.; and Gedeon, T. 2021b. Invertible Denoising Network: A Light Solution for Real Noise Removal. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 13365–13374.

Lore, K. G.; et al. 2017. LLNet: A deep autoencoder approach to natural low-light image enhancement. *Pattern Recognition*, 61: 650–662.

Lugmayr, A.; Danelljan, M.; Van Gool, L.; and Timofte, R. 2020. Srflow: Learning the super-resolution space with normalizing flow. In *European Conference on Computer Vision*, 715–732. Springer.

Lv, F.; Li, Y.; and Lu, F. 2021. Attention guided low-light image enhancement with a large scale low-light simulation dataset. *International Journal of Computer Vision*, 1–19.

Lv, F.; Lu, F.; Wu, J.; and Lim, C. 2018. MBLLEN: Low-Light Image/Video Enhancement Using CNNs. In *BMVC*, 220.

Pumarola, A.; Popov, S.; Moreno-Noguer, F.; and Ferrari, V. 2020. C-flow: Conditional generative flow models for images and 3d point clouds. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7949–7958.

Ren, W.; Liu, S.; Ma, L.; Xu, Q.; Xu, X.; Cao, X.; Du, J.; and Yang, M.-H. 2019. Low-light image enhancement via a deep hybrid network. *IEEE Transactions on Image Processing*, 28(9): 4364–4375.

Ren, X.; Li, M.; Cheng, W.-H.; and Liu, J. 2018. Joint enhancement and denoising method via sequential decomposition. In *2018 IEEE International Symposium on Circuits and Systems (ISCAS)*, 1–5. IEEE.

Risheng, L.; Long, M.; Jiaao, Z.; Xin, F.; and Zhongxuan, L. 2021. Retinex-inspired Unrolling with Cooperative Prior Architecture Search for Low-light Image Enhancement. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

Shen, L.; Yue, Z.; Feng, F.; Chen, Q.; Liu, S.; and Ma, J. 2017. Msr-net: Low-light image enhancement using deep convolutional network. *arXiv preprint arXiv:1711.02488*.

Tao, L.; Zhu, C.; Xiang, G.; Li, Y.; Jia, H.; and Xie, X. 2017. LLCNN: A convolutional neural network for low-light image enhancement. In *2017 IEEE Visual Communications and Image Processing (VCIP)*, 1–4.

Trippe, B. L.; and Turner, R. E. 2018. Conditional density estimation with bayesian normalising flows. *arXiv preprint arXiv:1802.04908*.

Wang, R.; Zhang, Q.; Fu, C.-W.; Shen, X.; Zheng, W.-S.; and Jia, J. 2019a. Underexposed photo enhancement using deep illumination estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6849–6857.

Wang, X.; Yu, K.; Wu, S.; Gu, J.; Liu, Y.; Dong, C.; Qiao, Y.; and Change Loy, C. 2018. Esrgan: Enhanced super-resolution generative adversarial networks. In *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*.

Wang, Y.; Cao, Y.; Zha, Z.-J.; Zhang, J.; Xiong, Z.; Zhang, W.; and Wu, F. 2019b. Progressive retinex: Mutually reinforced illumination-noise perception network for low-light image enhancement. In *Proceedings of the 27th ACM International Conference on Multimedia*, 2015–2023.

Wang, Z.; Bovik, A. C.; Sheikh, H. R.; and Simoncelli, E. P. 2004. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4): 600–612.

Wei, C.; Wang, W.; Yang, W.; and Liu, J. 2018. Deep retinex decomposition for low-light enhancement. *arXiv preprint arXiv:1808.04560*.

Winkler, C.; Worrall, D.; Hoogeboom, E.; and Welling, M. 2019. Learning likelihoods with conditional normalizing flows. *arXiv preprint arXiv:1912.00042*.

Wolf, V.; Lugmayr, A.; Danelljan, M.; Van Gool, L.; and Timofte, R. 2021. Deflow: Learning complex image degradations from unpaired data with conditional flows. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 94–103.

Xiao, M.; Zheng, S.; Liu, C.; Wang, Y.; He, D.; Ke, G.; Bian, J.; Lin, Z.; and Liu, T.-Y. 2020. Invertible image rescaling. In *European Conference on Computer Vision*, 126–144. Springer.

Yang, G.; Huang, X.; Hao, Z.; Liu, M.-Y.; Belongie, S.; and Hariharan, B. 2019. Pointflow: 3d point cloud generation with continuous normalizing flows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 4541–4550.

Yang, W.; Wang, S.; Fang, Y.; Wang, Y.; and Liu, J. 2020. From fidelity to perceptual quality: A semi-supervised approach for low-light image enhancement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3063–3072.

Ying, Z.; Li, G.; and Gao, W. 2017. A bio-inspired multi-exposure fusion framework for low-light image enhancement. *arXiv preprint arXiv:1711.00591*.

Zamir, S. W.; Arora, A.; Khan, S.; Hayat, M.; Khan, F. S.; Yang, M.-H.; and Shao, L. 2020. Learning enriched features for real image restoration and enhancement. *arXiv preprint arXiv:2003.06792*.

Zhang, R.; Isola, P.; Efros, A. A.; Shechtman, E.; and Wang, O. 2018. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 586–595.

Zhang, Y.; Guo, X.; Ma, J.; Liu, W.; and Zhang, J. 2021. Beyond Brightening Low-light Images. *International Journal of Computer Vision*, 129(4): 1013–1037.

Zhang, Y.; et al. 2019. Kindling the darkness: A practical low-light image enhancer. In *Proceedings of the 27th ACM International Conference on Multimedia*, 1632–1640.