ReCoRo: Region-Controllable Robust Light Enhancement with User-Specified Imprecise Masks

Dejia Xu*
University of Texas at Austin
USA

Hayk Poghosyan*
Picsart AI Research
Armenia

Shant Navasardyan
Picsart AI Research
Armenia

Yifan Jiang
University of Texas at Austin
USA

Humphrey Shi†
Picsart AI Research, UO & UIUC
USA

Zhangyang Wang†
University of Texas at Austin
USA

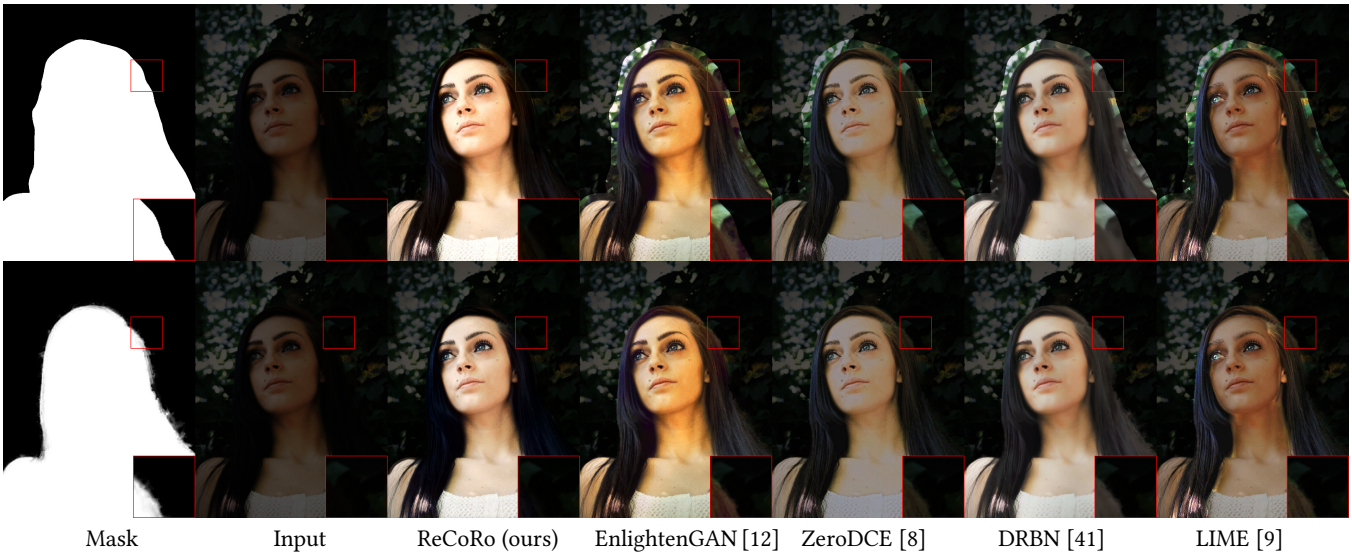


Figure 1: Our proposed ReCoRo is able to provide Region-Controllable Robust light enhancement with both user-specified imprecise and fine matting masks. Since no previous method is capable of such a controlling mechanism, we have to combine alpha blending with these global enhancement methods as baselines. They fail to perform well with alpha compositing on the imprecise masks (first row). For precise masks, they may suffer from clearly visible illumination level transitions (second row).

ABSTRACT

Low-light enhancement is an increasingly important function in image editing and visual creation. Most existing enhancing algorithms are trained to enlighten a given image in a globally homogeneous way, and (implicitly) to some predefined extent of brightness. They are neither capable of enhancing only local regions of interest (*"where"*) while keeping the overall visual appearance plausible, nor producing outputs at a range of different illumination levels (*"how much"*). Those hurdles significantly limit the prospect of flexible, customizable, or even user-interactive low-light enhancement. To address these gaps, we propose *Region-Controllable Robust Light Enhancement* (**ReCoRo**), a novel framework that allows users to directly specify "where" and "how much" they want to enhance from an input low-light image; meanwhile, the model will learn to intelligently maintain the overall consistent visual appearance and plausible composition via a discriminator. Moreover, since in practical mobile APPs, such user specifications often come in imprecise forms (e.g., finger-drawn masks), we propose to bake in domain-specific data augmentations into training ReCoRo, so that the learned model can gain resilience to various roughly-supplied user masks. Up to our best knowledge, ReCoRo is the first of its kind that allows the user to localize the enlightenment region as well as to control the light intensity. Extensive experiments clearly demonstrate that ReCoRo outperforms state-of-the-art methods in terms of qualitative results, quantitative metrics, and versatile controllability. Project repository: https://bit.ly/ReCoRo-lowlight.

*Both authors contributed equally to this research.
†Co-corresponding authors.

MM '22, October 10–14, 2022, Lisboa, Portugal.

ACM ISBN 978-1-4503-9203-7/22/10…$15.00
https://doi.org/10.1145/3503161.3547813

CCS CONCEPTS

• **Computing methodologies** → **Computer graphics**; **Computer vision**.

KEYWORDS

neural networks, image processing, light enhancement

ACM Reference Format:
Dejia Xu, Hayk Poghosyan, Shant Navasardyan, Yifan Jiang, Humphrey Shi, and Zhangyang Wang. 2022. ReCoRo: Region-Controllable Robust Light Enhancement with User-Specified Imprecise Masks. In *Proceedings of the 30th ACM International Conference on Multimedia (MM '22), October 10–14, 2022, Lisboa, Portugal.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3503161.3547813

1 INTRODUCTION

The low-light condition of images arises from a wide range of factors, including environmental cases like low illumination or more technical ones like sub-optimal ISO settings. Affected images are generally considered less pleasant and can introduce new challenges for downstream tasks as well. The rectification of such degradation has been an important research direction for many years, resulting in a wide variety of low-light enhancement solutions and approaches.

Low-light image enhancement is an ill-posed problem with a lack of large paired publicly available datasets, making it quite challenging to create a universal solution for 'one-click' enlightenment of images. In the past decades, numerous researchers have attempted to address this problem. For example, histogram equalization (HE) methods [1, 29] stretch the dynamic range of the low-light images yet tend to produce undesirable illumination in complex scenes.

Another line of research adopts the Retinex theory [17] to decompose a given low-light image into two layers, reflectance, and illumination, separately. Several filters [13, 20, 39] and hand-crafted priors [6, 9] are adopted to improve the decomposition.

With the rapid development of deep learning, new learning-based approaches for low-light enhancement have gained much interest recently. Various research has been proposed to restore the normal light images from the complex degradation with the help of large-scale datasets. Most methods [8, 31, 34, 38, 40] require well-aligned paired low-light and normal-light images to obtain effective performance. EnlightenGAN [12] makes the first effort to train the model without paired supervision. DRBN [41] further proposes a semi-supervised learning framework for low-light image enhancement, benefiting from both paired and unpaired data.

However, since they are designed to enlighten input images in a globally homogeneous way, most existing approaches usually struggle to handle extreme cases, such as back-lit images where both over-exposed and under-exposed regions are present. Additionally, when a user wants to enhance a specific region of the image, previous approaches would have to first enhance the whole image and then use alpha blending to obtain the final results. However, this approach produces a hard-coded transition region on the image and is often unrealistic in complex scenes. Moreover, this design requires the masks to be precise, which is not always the case for end-users on smartphones. A user-specified mask tends to be imprecise due to unprecedented noise, such as finger shaking. In these cases, existing methods produce sharp boundaries strictly aligned with the roughly-supplied masks, as shown in Fig. 1. Alleviating the need for precise masks from users is a valuable functionality in need. On the other hand, most previous methods implicitly learn to enlighten images to some predefined extent of brightness as they extract knowledge from paired datasets. On the contrary, the desired brightness level of a real-world image is highly subjective, and providing users with the flexibility to tune the brightness level of images is greatly needed.

To address these issues we propose **ReCoRo**, a novel framework that integrates a convenient and intuitive control mechanism, allowing the user to directly specify the region ("where") and the target illumination level ("how much") they want to enhance from an input low-light image. We propose a masked generative adversarial model, which integrates the controlling masks from the user via SPADE blocks [27]. We use the nonzero regions inside the mask to specify "where" to enhance and consider the values of the mask as the signal of "how much" to enhance. With the help of a local and a global discriminator, our model learns to generate an overall consistent visual appearance and plausible composition of the enhanced regions and the unmasked regions. We further tackle the real-world challenging cases in practical mobile APPs, by considering the input noises from imprecise user specifications. End users usually draw masks with fingers, thus the input masks consist of distorted boundaries with jagged shapes. In order to help the model gain resilience against these roughly-supplied user masks, we propose to bake in domain-specific data augmentations into the training process.

Our contributions can be summarized as follows:

- Up to our best knowledge, ReCoRo is the first attempt to allow users to enlighten localized regions and control the target light intensity as well.
- We further take the imprecise user-specified masks into consideration and propose domain-specific data augmentations to enable the network to gain resilience against the roughly-provided user masks.
- Extensive experiments clearly demonstrate our superiority against state-of-the-art methods including quantitative, qualitative, and in terms of versatile controllability as well.

2 RELATED WORKS

2.1 Traditional Methods

Traditional methods explore various image priors for single image low-light enhancement. Approaches based on local and global histogram equalization [1, 29] address the issue from a contrast increase angle. Other solutions [21, 44] consider the low-light images as inverted haze images and use dehazing methods for low-light enhancement. Another popular class of approaches is based on the Retinex theory [17]. Such methods decompose the image into two layers, illumination and reflectance, after which simple transformations are used to get the desired effects. SRIE [5] estimates both layers simultaneously with a weighted variational model. LIME [9], however, only estimates illumination and uses the reflection layer as the final enhanced results. JED [32] further improves the results by considering noise suppression via sequential decomposition. Although these methods produce promising results in some cases,

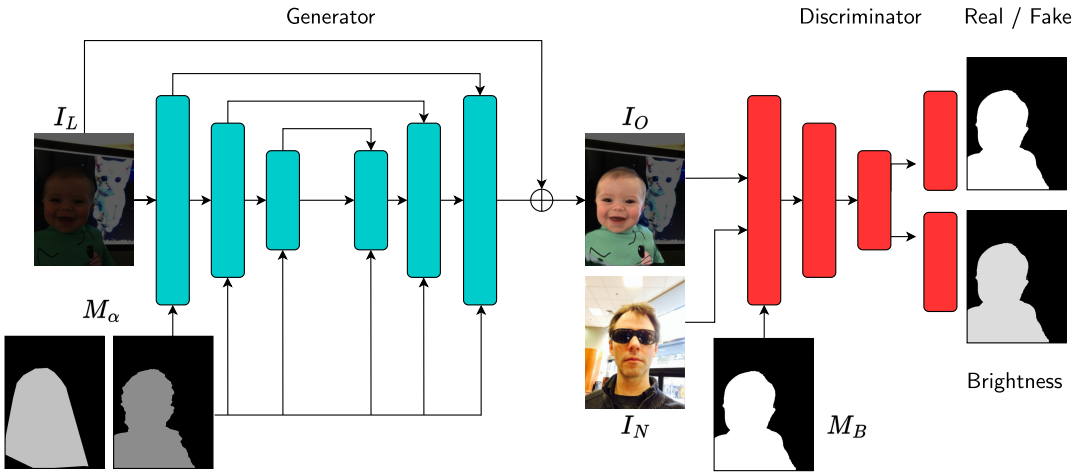


Figure 2: Overview of our proposed ReCoRo. In each training iteration, an input image I_L, together with a randomly generated rough mask M_α, is sent into the generator. M_α is integrated into the network in each of the layers using SPADE [27] blocks. The discriminators not only learn to distinguish whether the images are real or not, but also estimate the brightness level of their input (I_O or I_N). Another discriminator that focuses on randomly cropped regions is omitted for simplicity.

the hand-crafted models rely on careful parameter tuning and are limited by the model capacity.

2.2 Data-Driven Methods

Existing approaches can be generally categorized by their architecture and the type of datasets used. Generating natural paired images has historically been a huge challenge for low-light enhancement. Established techniques include fixing a camera and reducing the exposure time in normal-light conditions or increasing exposure time in low-light conditions [3]. These are difficult to scale and have their own limitations like capturing non-stationary scenes. LLNet and S-LLNet introduced in [23] are deep autoencoder-based approaches with the objectives of contrast-enhancement and denoising, they are trained on synthetically darkened data simulated with random Gamma corrections and additional noise. Based on the Retinex theory the similarly named Retinex-Net [3] postulates that paired images have the same reflectance but different illumination. The model is trained with the objective of lightness enhancement and denoising on the paired LOw-Light (LOL) data set, collected by the authors. Since there are no large paired datasets and previous methods don't generalize well, much effort has been made to improve the performance without the need for paired supervision. EnlightenGAN [12] based on the generative adversarial network architecture is the first to succeed in using unpaired data. Zero-DCE [8] formulates light enhancement as a task of image-specific curve estimation with a deep network. CERL [4] improves upon EnlightenGAN and introduces plug-and-play noise suppression. In [41], the proposed DRBN (deep recursive band network) model uses a novel band deconstruction/reconstruction architecture. Depending on different image scenarios (e.g. presence of extreme light conditions or rich semantics) the existing approaches may vary in terms of quality but generally assume that the images should be enlightened to a pre-defined brightness level, and avoid handling the real-world challenging case of region controlling. In contrast, our ReCoRo offers both visually pleasing results and controllability with the masked generative adversarial framework, and further handles roughly-supplied masks in real-world scenarios.

2.3 Adversarial Learning

Generative Adversarial Networks (GANs) have shown great success in image synthesis and image-to-image translation. Since the seminal work by Goodfellow *et.al* [7], numerous research has been conducted on improving the GAN models [2, 22, 25]. As GAN prove itself with a powerful ability to generate photo-realistic images [14–16], they are also adopted for image restoration tasks, including super resolution [18], noise removal, image inpainting [43], image colorization [11], rain removal [30], and low-light enhancement [12]. One line of research follows CycleGAN [45], which adopts two generators to translate between two different domains. A cycle-consistency loss is introduced on unpaired data. Another direction removes the need for cycle consistency and uses a one-path design. EnlightenGAN [12] adopts semi-supervised learning to utilize both paired and unpaired training sets. CUT [26] leverages contrastive learning operating on patches of the original image and enables one-sided translation in the unpaired image-to-image translation setting. The one-path design is more favored since it is more computationally efficient. ReCoRo also adopts a one-path design so that the network is lightweight and stable.

3 METHOD

3.1 Overview

The setting of region-controllable robust light enhancement is challenging, as there are no available paired real-world images for training. To tackle this problem, we build our ReCoRo framework as a semi-supervised framework to train on both paired synthetic and unpaired real-world images. The paired synthetic dataset is considered as a labeled set, while the real-world low-light and

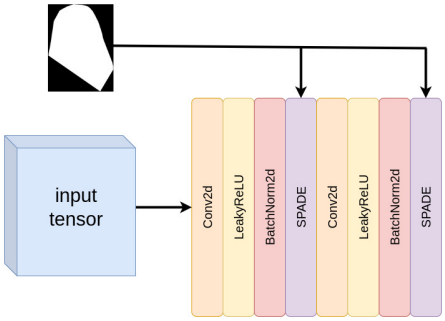


Figure 3: The base block of our generator.

normal-light images are taken as an unlabeled set. ReCoRo controls the desired enhancement intensity with the α map input. We associate a zero value in the α map with the preference of no enhancement at all, and the value of one with the enhancement to a well-illuminated pixel. For global enhancement, the α map is a constant map. Supervision via discriminator and self-regularization loss is implemented on both the labeled set and the unlabeled set, while pixel-wise alignment is additionally enforced on the labeled set.

In the following sections, we first presume the masks are precise and introduce the masked generative adversarial network that provides the region-control functionality for users. Then, we tackle the challenging real-world cases when the user masks are roughly-supplied. We illustrate the design of our domain-specific mask augmentations that enable the network to be robust against these roughly-supplied masks.

3.2 Masked Generator

As shown in Fig. 2, our generator adopts a U-Net [33] architecture. We implement the region-control functionality using SPADE layers [27], and plug in the SPADE layer after every other convolution layer. Different from the common usage of using the SPADE layer for semantic region control [27], we further extend the functionality of the masks by linearly scaling them with an additional controlling parameter α. In summary, the non-zero regions of the masks decide "where" to enhance and the exact value of the control of the mask "how much" the enhancement shall take place.

The architecture of the base block of our generator is provided in Fig. 3. After each convolution layer, we adopt a LeakyReLU layer as activation. The activated features are then passed into a BatchNorm layer [10]. Finally, we integrate the normalized features with the user-specified masks via the SPADE blocks.

Inspired by EnlightenGAN [12], we adopt a simple yet effective self-regularization design. Since the ImageNet-trained classification models are robust against intensity change of the input images [12], the low-light images and their normal-light variants are close in terms of distance in VGG-16 [36] feature space. We enforce a self-feature preserving loss $\mathcal{L}_{\text{SFP}}$ for self-regularization:

$$\mathcal{L}_{\text{SFP}}(I_L) = \frac{1}{W_{i,j}H_{i,j}} \sum_{x=1}^{W_{i,j}} \sum_{y=1}^{H_{i,j}} (\phi_{i,j}(I_L) - \phi_{i,j}(I_O)))^2, \tag{1}$$

where I^L is the input low-light image, I_O refers to the output enhanced image, and $\phi_{i,j}$ denotes the j-th convolution layer after the i-th max-pooling layer of the pre-trained VGG-16 model. For all our experiments, we empirically set $i = 5$ and $j = 1$.

3.3 Masked Discriminator

We adopt two discriminators in our design, one for global fidelity and the other for local structure. However, the vanilla incorporation of a global and a local discriminator as in EnlightenGAN [12] is not expressive enough to validate the quality of the output images under diverse regions and adaptive α control. Thus, we propose to integrate the masking mechanism into the discriminators.

The global discriminator focuses on validating the realism of the whole image. It is asked to determine whether the overall image is natural or not and takes the whole output image I_O of the generator as input. On the other hand, the local discriminator only sees some local regions of the images and focuses on the quality of these patches. They are both also fed with the masks M_B so that they can validate whether the output images align with the masks. The architecture of the discriminators is a sequence of convolution and LeakyReLU layers. We further adopt an auxiliary branch for each discriminator to estimate the brightness level of the input images I_O. This auxiliary branch pushes the discriminators to not only focus on the quality of the images but also take the intensity values into consideration:

$$\mathcal{L}_{\text{AUX}} = ||D_{\text{aux}}(I_O) - M_p||, \tag{2}$$

where D_{aux} refers to the output of the auxiliary branch of the discriminators and M_p is the ground truth brightness level for I_O. In this way, the discriminators are encouraged to understand the inherent contents of the image while being aware of the brightness level. Note that the input mask M_B for discriminators is a binarized precise mask and the discriminator has to estimate M_p which inherently contains the controlling α.

Both discriminators are Markovian Discriminators, and the loss functions are formulated as follows,

$$\begin{aligned} L_D &= \mathbb{E}_{\boldsymbol{I_N} \sim p(\boldsymbol{I_N})} \left[f_D(-D(\boldsymbol{I_N})) \right] + \mathbb{E}_{\boldsymbol{I_L} \sim p(\boldsymbol{I_L})} \left[f_D(D(G(\boldsymbol{I_L}))) \right] \\ L_G &= \mathbb{E}_{\boldsymbol{I_L} \sim p(\boldsymbol{I_L})} \left[f_G(-D(G(\boldsymbol{I_L}))) \right]. \end{aligned} \tag{3}$$

where G and D refers to the generator and discriminator, respectively. We train the GAN framework using Hinge Loss [22], so $f_D(x) = \max(0, 1 + x)$ and $f_G(x) = x$.

3.4 Domain-specific Mask Augmentations

The aforementioned design is able to obtain fairly effective performance when our model is allowed to rely on the input masks. In such an ideal laboratory setting, our model is capable of generating high-quality results that align well with the input masks, as shown in Sec. 4.3. Real-world cases, however, contain much noisier input and the masks are often distorted. It is common, for example, that an end-user draws the masks using fingers on a mobile app. Such application leads to the demand for the model to be able to handle imprecise masks.

The lack of objective metrics and paired annotations makes obtaining a visually-pleasing result from the imprecise user masks

Figure 4: Illustration of our domain-specific mask augmentations. Rows from left to right: the original matting mask, coarse, wave distorted, dilated, and eroded augmentations.

essentially difficult. We tackle this problem by introducing several domain-specific mask augmentations during the training process. By augmenting the user masks, our ReCoRo is capable of learning invariant enhancement results even when the masks are roughly specified.

We empirically adopt four kinds of augmented masks: 1) wave distorted masks; 2) coarse boundary masks; 3) dilated masks; 4) eroded masks. As shown in Fig. 4, the augmented masks are roughly the same as the precise masks, while being slightly different in the boundary regions. Such design aligns well with the real-world user-specified masks, where users tend to cover the majority of the target region with masks but are often too impatient to refine the mask boundaries.

In each training iteration on the human matting dataset, given a certain input image, we randomly sample a rough mask and provide it to the model alongside the precise mask. As shown in Fig. 2, the generator takes the rough mask M_α as input and is forced to produce a plausible composition aligned with the precise mask M_p. Note that under the rough mask setting, the input of the discriminator is not as distorted as M_α, but are binarized precise masks. In summary, the discriminator always takes the precise masks as input, while the generator is fed with different versions of masks and is forced to produce invariant results.

We also introduce a mask-guided loss using the precise masks, regardless of which coarse mask is selected as input for the generator. A mask penalizing loss enforces the unmasked regions to be left untouched:

$$\mathcal{L}_P = ||(1 - M_B) * (I_O - I_L)||_2, \tag{4}$$

The above-mentioned loss functions are both calculated on the labeled and unlabeled sets. The only difference between the loss functions on the two datasets is that we also enforce pixel-wise alignment on the labeled data:

$$\mathcal{L}_{\text{PIX}} = ||I_O - I_N||_2. \tag{5}$$

The overall loss function is summarized as follows:

$$\mathcal{L} = \mathcal{L}_P + \lambda_1(\mathcal{L}_D + \mathcal{L}_G) + \lambda_2\mathcal{L}_{\text{SFP}} + \lambda_3\mathcal{L}_{\text{AUX}} + \lambda_4\mathcal{L}_{\text{PIX}}, \tag{6}$$

Model/Metric	SSIM↑	PSNR↑	LPIPS↓
EnlightenGAN	0.7510	17.3139	0.2241
RetinexNet	0.5821	14.5458	0.3921
ZeroDCE	0.7348	17.6303	0.2189
DRBN	**0.7829**	16.6195	0.1994
ReCoRo (ours, $\alpha = 1$)	0.7661	**17.7148**	**0.1932**

Table 1: Quantitative comparison on the LOL dataset.

where λ_1, λ_2, λ_3 and λ_4 are weighting factors. Note that for each training iteration, the loss functions are calculated not only on the whole input image but also on several random cropped patches because of the local discriminator.

4 EXPERIMENTS

More experiments and analyses are provided in the supplementary material.

4.1 Dataset

The data used for the training process consists of two distinct subsets. The first is identical to the dataset used in EnlightenGAN [12]. It is an agglomerated dataset collected from datasets used in previous works [9, 19, 24, 37, 39]. It includes 914 low light and 1016 normal light images. The second is a paired dataset synthesized from the Human Matting dataset [35]. It has 1,700 images collected from Flickr. The images cover a good variety of ages, colors, clothing, accessories, hairstyle, head position, background scene, etc. All images are cropped such that the face rectangles are of similar sizes. The synthesis of low-light images is done by blending the Gamma corrected and HSL lightness channel scaled versions of the normal light image. This dataset originally includes pixel-wise matting masks, and we randomly generate 4 kinds of coarse versions (wave distorted, coarse boundary, dilated, and eroded, as illustrated in Sec. 3.4) for each precise mask.

The test set is also two-fold. For the LOL dataset, we use the same partition as in EnlightenGAN, which consists of 148 paired images. For the human matting dataset, there are 300 low-light and normal-light pairs.

4.2 Evaluation Protocol

We compare our method against state-of-the-art "one-click" low-light enhancement algorithms. Deep learning methods include EnlightenGAN [12], RetinexNet [40], ZeroDCE [8], and DRBN [41]. As training datasets are often the contributions of existing methods, so we obtain the results of these comparison methods using their released pre-trained models. We also compare with LIME [9], which is a traditional method based on the Retinex theory.

4.3 Local Enhancement

Qualitative Comparison. We provide visual comparisons in Fig. 5. Our ReCoRo achieves superior qualitative results to previous methods. In general, the previous methods suffer from inconsistent boundaries and harsh transition regions. These global enhancement methods generate results with color distortions or undesired

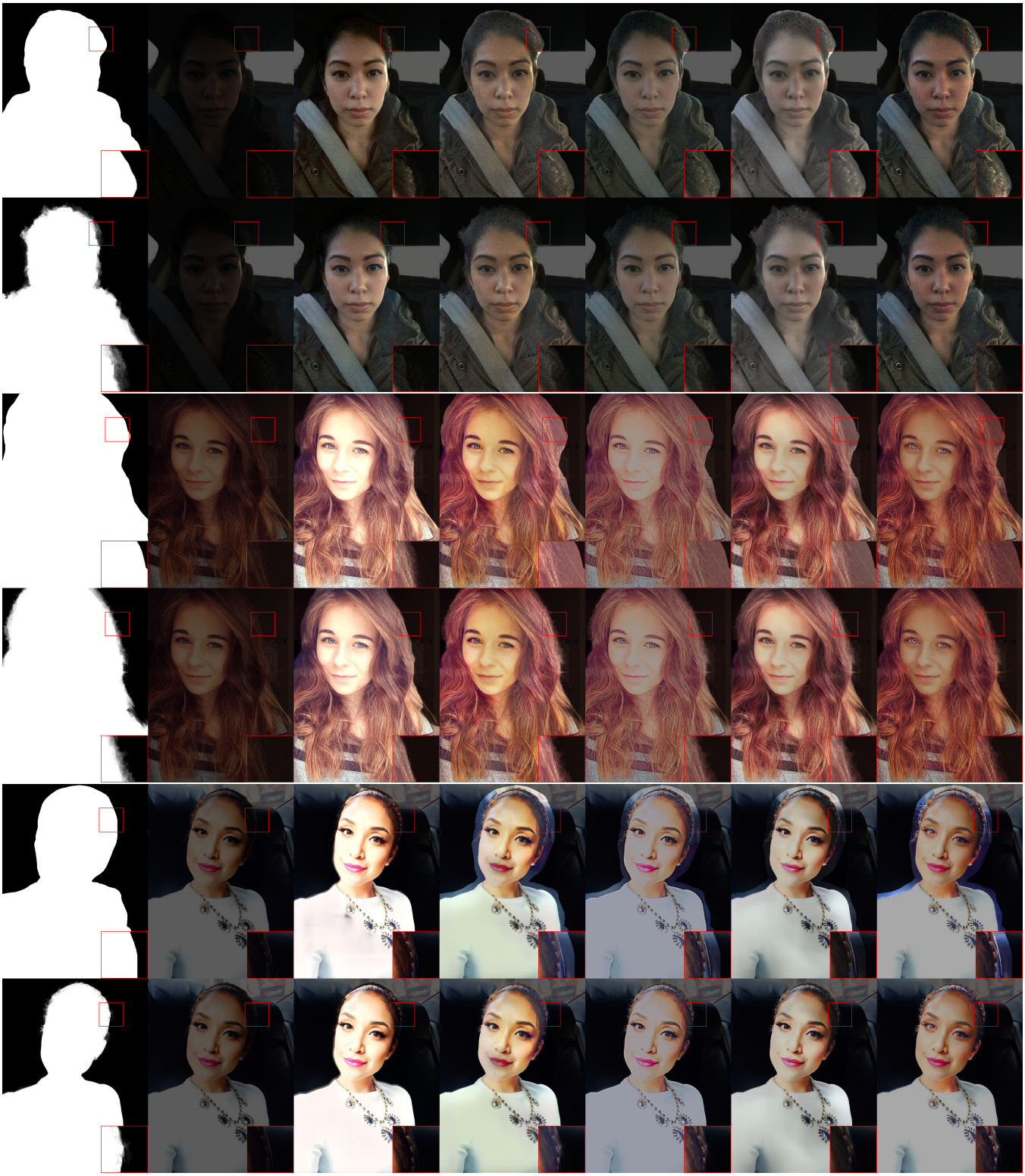
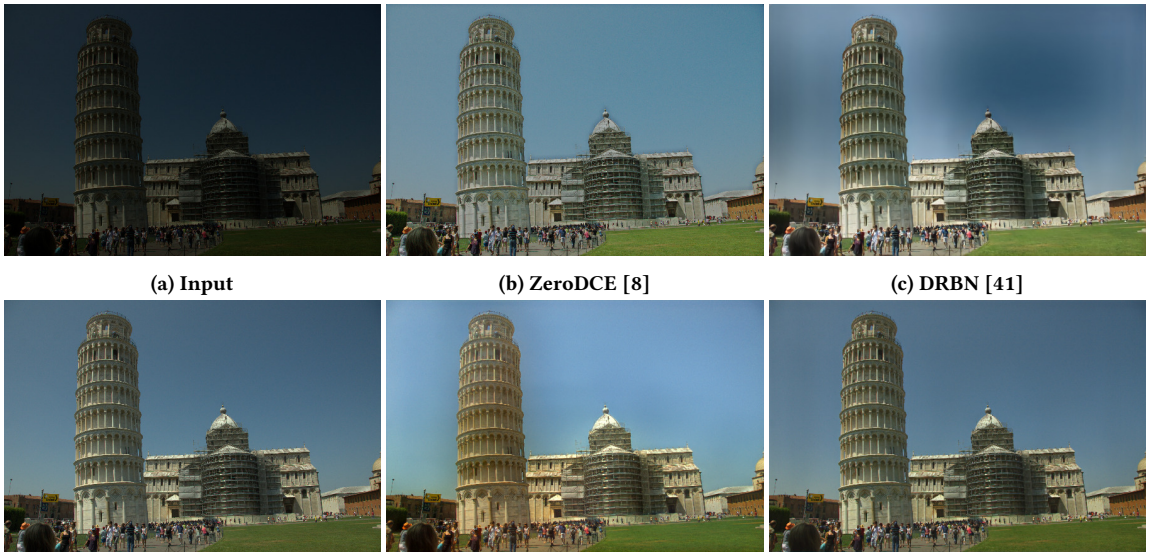


Figure 5: Visual comparison on region controllable enhancement. Each set contains two rows: the first row shows results when the input masks are roughly supplied; the second row shows the results for precise masks. As previous methods are not specially designed for local enhancement, we use alpha blending to obtain their results. Our method is capable of generating plausible composition and visually-pleasing enhancement results, even when the model is provided with imprecise masks.



(a) Input (b) ZeroDCE [8] (c) DRBN [41]

(d) Target Image (e) EnlightenGAN [12] (f) ReCoRo (ours), $\alpha = 1$

Figure 6: Visual comparisons of global light enhancement on the LOL dataset.

Input $\alpha = 0$ $\alpha = 0.5$ $\alpha = 1$ Normal Light

Figure 7: Visual results of our method with different α for global enhancement. Each row corresponds to the same input image. Each column shows results from the same masks.

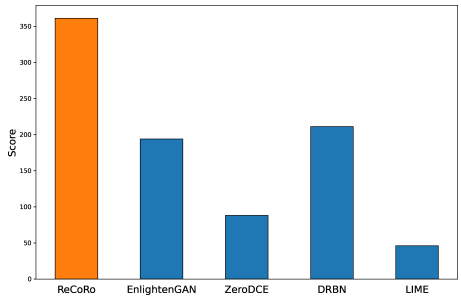


Figure 8: Human subjective study of region-controllable robust light enhancement. The x-axis lists different methods and the y-axis shows the scores.

brightness levels. When the comparison methods are provided with rough masks, due to the limitation they can only obtain unnatural results strictly aligned with the input masks. One can notice that for example, when the masks mistakenly have jagged boundaries, the enhanced results also contain such artifacts in the transition regions. In contrast, our method achieves very good perceptual quality, with plausible transition regions, even when provided with roughly-specified masks.

Human Subjective Study. In the absence of comparable methods and data for the local enhancement, one of the few ways available to validate our results is based on the subjective preference of humans. We conduct a human subjective study to compare ReCoRo with four other approaches: EnlightenGAN, ZeroDCE, DRBN, and LIME. To obtain images corresponding to local α mask enhancement from other approaches we alpha blend their enhanced images

Model/Metric	SSIM↑	PSNR↑	LPIPS↓
w/o mask augmentations	0.7593	17.2177	0.1989
w/o local discriminator	0.6538	15.0091	0.3681
w/o semi supervised training	0.7395	**17.8922**	0.2209
ReCoRo	**0.7661**	17.7148	**0.1932**

Table 2: Ablation study quantitative comparison with $\alpha = 1$ evaluated on the LOL dataset.

Figure 9: Visual comparison of our ablation studies. The first row shows images of the low light input image, imprecise mask, and full model (ReCoRo), respectively. The second row shows the results of 3 variants: w/o mask augmentations, w/o semi-supervised training, and w/o local discriminator.

with the low light input image using the user-specified mask. We randomly select 30 images from our paired person dataset and use annotations similar to what a non-professional user would use i.e. the annotations were done focusing on overall speed instead of the quality of details. We randomly shuffle the results of five approaches(one native and four blends) and ask 5 subjects to choose the best three results in ascending order from best to worst. The subjects were instructed to compare using the criteria of visual appeal and in accordance with the expected results. The results show a clear preference for images enhanced by ReCoRo see Fig. 8. The resulting score is a weighted sum where for each image each model is allocated 3 points if it's the best 2 if it's the second-best and 1 if it's third-best. Out of 150 cases ReCoRo was the most preferable in 101 cases($\approx$ 67%), the second most preferable in 24 cases($\approx$ 16%) and third most preferable in 10 cases($\approx$ 6%).

4.4 Global Enlightening

Qualitative Comparison. The vanilla global enlightening task can be considered a special case of our region-controllable enhancement setting. Hence the input mask contains a constant and covers the whole image, the model is able to generate homogeneous enlightened results, as shown in Fig. 6. One can see that our method generates the most visually-pleasing results with neither color distortion nor ghosting effect in flat regions. On the contrary, ZeroDCE suffers from incorrect color and DRBN and EnlightenGAN demonstrate ghosting shadows undesirably.

Quantitative Comparison. Table 1 shows quantitative comparisons against state-of-the-art global enhancement methods on the LOL dataset. We evaluate our model using a constant α map of value one, which corresponds to well-illuminated results. ReCoRo outperforms almost all state-of-the-art methods, with the exception of DRBN on the SSIM metric. This is partially explained by the broad functionality of our approach, and the fact that DRBN uses the LOL dataset [40] and its extended version [42] in a paired setting.

Controlling Functionality. We also demonstrate the controlling functionality of the α map for ReCoRo in Fig. 7. Given the same input image with various α map control, our ReCoRo is capable of generating different outputs with different brightness levels in a visually-pleasing way. Note that for $\alpha = 0$, our model leaves the input images untouched. And for $\alpha = 1$, the images are enhanced to a well-illuminated brightness level.

4.5 Ablation Study

We perform ablation studies to justify the effectiveness of our core components: the domain-specific mask augmentation, the local discriminator, and the semi-supervised training scheme. In all these cases we observe a visible decrease either in $\alpha = 1$ metrics(shown in Tab. 2) or visual α map enhancements (shown in Fig 9). ReCoRo w/o mask augmentations show significantly less mask adjustment resulting in visually unappealing white edges. The variant without the local discriminator has much lower metrics. And the model without the semi-supervised training scheme suffers from less detail in the output images since this variant utilizes less training data.

5 CONCLUSIONS

We present ReCoRo, a novel framework for users to directly specify "where" and "how much" they want to enhance from an input low-light image. Meanwhile, the model learns to intelligently maintain the overall consistent visual appearance and plausible composition via a discriminator. We further tackle the real-world scenarios where user-specified masks are roughly-supplied and the model gains resilience via domain-specific data augmentations during the training process. Comprehensive experiments are conducted on various datasets, where ReCoRo outperforms existing light enhancement methods.

REFERENCES

[1] M. Abdullah-Al-Wadud, M. H. Kabir, M. A. Akber Dewan, and O. Chae. 2007. A Dynamic Histogram Equalization for Image Contrast Enhancement. *IEEE Transactions on Consumer Electronics* 53, 2 (May 2007), 593–600.

[2] Martin Arjovsky, Soumith Chintala, and Léon Bottou. 2017. Wasserstein generative adversarial networks. In *International conference on machine learning*. PMLR, 214–223.

[3] Wei Chen, Wang Wenjing, Yang Wenhan, and Liu Jiaying. 2018. Deep Retinex Decomposition for Low-Light Enhancement. In *British Machine Vision Conference*. British Machine Vision Association.

[4] Zeyuan Chen, Yifan Jiang, Dong Liu, and Zhangyang Wang. 2022. CERL: A Unified Optimization Framework for Light Enhancement With Realistic Noise. *IEEE Transactions on Image Processing* (2022).

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

[6] X. Fu, D. Zeng, Y. Huang, X. P. Zhang, and X. Ding. 2016. A Weighted Variational Model for Simultaneous Reflectance and Illumination Estimation. In *Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition*. 2782–2790.

[7] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial nets. *Advances in neural information processing systems* 27 (2014).

[8] Chunle Guo, Chongyi Li, Jichang Guo, Chen Change Loy, Junhui Hou, Sam Kwong, and Runmin Cong. 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.

[9] X. Guo, Y. Li, and H. Ling. 2017. LIME: Low-Light Image Enhancement via Illumination Map Estimation. *IEEE Trans. on Image Processing* 26, 2 (Feb 2017), 982–993.

[10] Sergey Ioffe and Christian Szegedy. 2015. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International conference on machine learning*. PMLR, 448–456.

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

[12] Yifan Jiang, Xinyu Gong, Ding Liu, Yu Cheng, Chen Fang, Xiaohui Shen, Jianchao Yang, Pan Zhou, and Zhangyang Wang. 2021. Enlightengan: Deep light enhancement without paired supervision. *IEEE Transactions on Image Processing* 30 (2021), 2340–2349.

[13] D. J. Jobson, Z. Rahman, and G. A. Woodell. 1997. A multiscale retinex for bridging the gap between color images and the human observation of scenes. *IEEE Trans. on Image Processing* 6, 7 (Jul 1997), 965–976.

[14] Tero Karras, Miika Aittala, Samuli Laine, Erik Härkönen, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. 2021. Alias-free generative adversarial networks. *Advances in Neural Information Processing Systems* 34 (2021).

[15] Tero Karras, Samuli Laine, and Timo Aila. 2019. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 4401–4410.

[16] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. 2020. Analyzing and improving the image quality of stylegan. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 8110–8119.

[17] Edwin H. Land. 1977. The retinex theory of color vision. *Sci. Amer* (1977), 108–128.

[18] Christian Ledig, Lucas Theis, Ferenc Huszár, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, et al. 2017. Photo-realistic single image super-resolution using a generative adversarial network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 4681–4690.

[19] C. Lee, C. Lee, and C. S. Kim. 2013. Contrast Enhancement Based on Layered Difference Representation of 2D Histograms. *IEEE Trans. on Image Processing* 22, 12 (Dec 2013), 5372–5384.

[20] Chang-Hsing Lee, Jau-Ling Shih, Cheng-Chang Lien, and Chin-Chuan Han. 2013. Adaptive multiscale retinex for image contrast enhancement. In *Signal-Image Technology & Internet-Based Systems (SITIS), 2013 International Conference on*. IEEE, 43–50.

[21] L. Li, R. Wang, W. Wang, and W. Gao. 2015. A low-light image enhancement method for both denoising and contrast enlarging. In *Proc. IEEE Int'l Conf. Image Processing*. 3730–3734.

[22] Jae Hyun Lim and Jong Chul Ye. 2017. Geometric gan. *arXiv preprint arXiv:1705.02894* (2017).

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

[24] K. Ma, K. Zeng, and Z. Wang. 2015. Perceptual Quality Assessment for Multi-Exposure Image Fusion. *IEEE Trans. on Image Processing* 24, 11 (Nov 2015), 3345–3356.

[25] Xudong Mao, Qing Li, Haoran Xie, Raymond YK Lau, Zhen Wang, and Stephen Paul Smolley. 2017. Least squares generative adversarial networks. In *Proceedings of the IEEE international conference on computer vision*. 2794–2802.

[26] Taesung Park, Alexei A Efros, Richard Zhang, and Jun-Yan Zhu. 2020. Contrastive learning for unpaired image-to-image translation. In *European Conference on Computer Vision*. Springer, 319–345.

[27] Taesung Park, Ming-Yu Liu, Ting-Chun Wang, and Jun-Yan Zhu. 2019. Semantic Image Synthesis with Spatially-Adaptive Normalization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

[28] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. 2017. Automatic Differentiation in PyTorch. In *NIPS Autodiff Workshop*.

[29] S. M. Pizer, R. E. Johnston, J. P. Ericksen, B. C. Yankaskas, and K. E. Muller. 1990. Contrast-limited adaptive histogram equalization: speed and effectiveness. In *Proceedings of Conference on Visualization in Biomedical Computing*. 337–345.

[30] Rui Qian, Robby T Tan, Wenhan Yang, Jiajun Su, and Jiaying Liu. 2018. Attentive generative adversarial network for raindrop removal from a single image. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2482–2491.

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

[32] Xutong Ren, Mading Li, Wen-Huang Cheng, and Jiaying Liu. 2018. Joint enhancement and denoising method via sequential decomposition. In *2018 IEEE international symposium on circuits and systems (ISCAS)*. IEEE, 1–5.

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

[34] Liang Shen, Zihan Yue, Fan Feng, Quan Chen, Shihao Liu, and Jie Ma. 2017. Msr-net: Low-light image enhancement using deep convolutional network. *arXiv preprint arXiv:1711.02488* (2017).

[35] Xiaoyong Shen, Xin Tao, Hongyun Gao, Chao Zhou, and Jiaya Jia. 2016. Deep automatic portrait matting. In *European conference on computer vision*. Springer, 92–107.

[36] Karen Simonyan and Andrew Zisserman. 2014. Very Deep Convolutional Networks for Large-Scale Image Recognition. *Proc. Int'l Conf. Learning Representations* (2014).

[37] Vassilios Vonikakis, Rigas Kouskouridas, and Antonios Gasteratos. 2018. On the Evaluation of Illumination Compensation Algorithms. *Multimedia Tools Appl.* 77, 8 (April 2018), 9211–9231.

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

[39] S. Wang, J. Zheng, H. M. Hu, and B. Li. 2013. Naturalness Preserved Enhancement Algorithm for Non-Uniform Illumination Images. *IEEE Trans. on Image Processing* 22, 9 (Sept 2013), 3538–3548.

[40] Chen Wei*, Wenjing Wang*, Wenhan Yang, and Jiaying Liu. 2018. Deep Retinex Decomposition for Low-Light Enhancement. In *British Machine Vision Conference*.

[41] Wenhan Yang, Shiqi Wang, Yuming Fang, Yue Wang, and Jiaying Liu. 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.

[42] Wenhan Yang, Wenjing Wang, Haofeng Huang, Shiqi Wang, and Jiaying Liu. 2021. Sparse Gradient Regularized Deep Retinex Network for Robust Low-Light Image Enhancement. *IEEE Transactions on Image Processing* 30 (2021), 2072–2086.

[43] Jiahui Yu, Zhe Lin, Jimei Yang, Xiaohui Shen, Xin Lu, and Thomas S Huang. 2018. Generative image inpainting with contextual attention. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 5505–5514.

[44] X. Zhang, P. Shen, L. Luo, L. Zhang, and J. Song. 2012. Enhancement and noise reduction of very low light level images. In *Proc. IEEE Int'l Conf. Pattern Recognition*. 2034–2037.

[45] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. 2017. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*. 2223–2232.

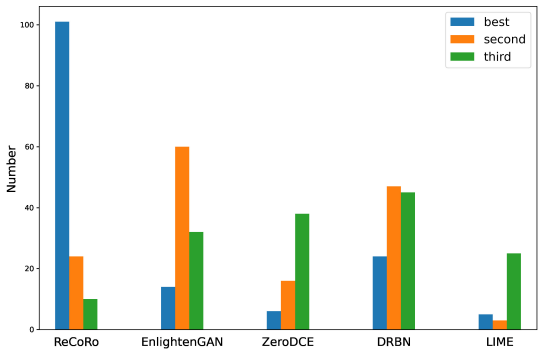


Figure 10: Human subjective study of region-controllable robust light enhancement. The x-axis lists different methods and the y-axis shows the number of subjects who chose this method.

A IMPLEMENTATION DETAILS

We implement our framework using PyTorch [28]. The data augmentation used during training includes transpose, horizontal, and vertical flipping. The optimization is done with Adam optimizer with a learning rate initialized to 1e-4 for the first 100 epochs and linearly decayed to 0 in the next 100 epochs. We use a batch size of 16 for all experiments. During training, we utilize patches of size 320×320 and randomly crop them to obtain 32×32 patches for the local discriminator. All experiments are done on an NVIDIA A6000 GPU.

B MORE COMPARISONS

B.1 Max similarity quantitative comparison

While measuring the global α enhancement performance we observe that the enhancement of other approaches results in an arbitrary amount of illumination which may or may not coincide with the corresponding real normal light image. In contrast to this, our method with its controllable property allows the user to choose an α and enhance the image in a way much more similar to the desired normal light image. To show this numerically we compute the following metric. For a paired low-light I_{low} and normal light I_{normal} images, one can compute the maximum similarity between ReCoRo enhanced images I_α(with a constant illumination map filled with the value $\alpha \in [0, 1]$) and I_{normal} and compare that with the similarity between the enhanced image $\hat{I}$ from our target comparison approaches. For a set of paired images $\{I_{i,low}, I_{i,normal} : i \in 1..n\}$, the maximum similarity for ReCoRo is

$$m_i = max\{M(I_{i,\alpha}; I_{i,normal}) : \alpha \in [0, 1]\} \tag{7}$$

where M is any similarity metric, the results of this comparison are shown in Table 3 The comparisons are done employing the same metrics and target models as in the first comparison.

Model/Metric	SSIM↑	PSNR↑	LPIPS↓
EnlightenGAN	0.7510	17.3139	0.2241
RetinexNet	0.5821	14.5458	0.3921
ZeroDCE	0.7348	17.6303	0.2189
DRBN	0.7829	16.6195	0.1994
ReCoRo	0.7661	17.7148	0.1932
max sim ReCoRo	**0.787**	**20.868**	**0.187**

Table 3: Max similarity quantitative comparison on the LOL dataset.

B.2 Global enhancement

We provide visual comparisons of global enhancement in Fig. 6 and Fig. 13.

B.3 Local enhancement

We provide visual comparisons of local enhancement for α maps with different values Fig. 14. In Fig 12 we show local enhancement visual comparisons for different augmented masks.

B.4 Human study

In the conducted human study each subject is given 30 images, where each contains the mask, low light input and five randomly shuffled enhancements (ReCoRo, EnlightenGAN, ZeroDCE, DRBN, LIME). In Fig 11 we show examples of images presented to the subjects. 10 shows the number of subjects preferring the indicated enhancement approach at a preference level. We also noted a high consensus level between our study subjects indicated that in 40.4% of cases, two subjects had the same level of preference for the enhanced image.

Figure 11: Examples presented to the human study subjects.

Figure 12: Example images with different augmentation masks. The images are positioned in a grid where the first column(y-axis) shows the low light inputs and the first row(x-axis) shows the used α maps, any other images in between are the enhanced versions with respect to the low light and map inputs.

Figure 13: Visual comparisons of global light enhancement on the LOL dataset.

Figure 14: Visual results of our method with different α maps(segmentation map * scalar) for local enhancement. At each row the images are: input low light image, enhanced image with $\alpha = 0$, enhanced image with $\alpha = 0.5$, enhanced image with $\alpha = 1$, real normal light image.