Pattern Recognition 131 (2022) 108867

Contents lists available at ScienceDirect

Pattern Recognition

journal homepage: www.elsevier.com/locate/patcog

Shedding light on images: Multi-level image brightness enhancement guided by arbitrary references

Ya'nan Wang[a,1], Zhuqing Jiang[a,b,1,*], Chang Liu[a], Kai Li[a], Aidong Men[a], Haiying Wang[a], Xiaobo Chen[c]

[a] School of Artificial Intelligence, Beijing University of Posts and Telecommunications, Beijing, China
[b] Beijing Key Laboratory of Network System and Network Culture, Beijing University of Posts and Telecommunications, Beijing, China
[c] Supervision Center, National Radio and Television Administration, Beijing, China

ARTICLE INFO

Article history:
Received 5 November 2021
Revised 12 June 2022
Accepted 19 June 2022
Available online 24 June 2022

Keywords:
Low-light image enhancement
Multi-level mapping
Arbitrary references
Codec network
Decomposition
Concatenation

ABSTRACT

The non-linearity between human perception and image brightness levels results in different definitions of NORMAL-light. Thus, most existing low-light image enhancement methods which produce one-to-one mapping can not meet the aesthetic demand. Other pioneers enhance low-light images guided by a given value. However, the inherent problem of non-linearity will cause poor usability. To this end, we propose a user-friendly neural network for multi-level low-light image enhancement. Inspired by style transfer, our method decomposes an image into content component feature and luminance component feature in the latent space. Then we enhance the image brightness to different levels by concatenating the content components from low-light images and the luminance components from reference images. The network meets various user requirements by selecting different brightness references. Moreover, information except for brightness is preserved to alleviate color distortion. Extensive experiments demonstrate the superiority of our network against existing methods.

1. Introduction

Presently, taking photos is convenient with the omnipresence of cameras on various devices. However, images often suffer degradation due to environmental and equipment limitations, i.e., low contrast, noise, and blur. Thus, image quality enhancement technologies have snowballed, including low-light image enhancement, dehazing [1], deblurring [2], deraining [3], image inpainting [4], and super-resolution [5]. Among them, low-light image enhancement is raising more and more attention, since an appropriate brightness is essential for both users feelings and downstream tasks. Since the NORMAL light level of images is justified by specific users, the image enhancement task should be guided by users aesthetics. Existing professional software help users get visually pleasing images by providing artificial photo manipulation tools. However, those tools are either user-unfriendly or with insufficient effect. Thus, it is essential to develop a flexible on-demand low-light image enhancement method that interacts with users and meets their preferences.

With the rapid development of deep learning, various approaches have been proposed to enhance low-light images. Some pioneers enhance low-light images to a fixed brightness. For instance, Lore et al. [6] enhances and denoises low-light images, Wei et al. [7] combines conventional theory with neural networks, Chen et al. [8] enhances raw images, and [9] utilizes a set of multi-scale features. Those algorithms learn the brightness difference of training data pairs, and enhance images without diversity, thus they are inflexible and ignore the user subjectivity. Some methods are proposed to tackle those problems. In [10], the light level is adjusted by a strength ratio. However, it may not be a wieldy descriptor for users, since the relationship between the perceived change of light level and the strength ratio is non-linear. Kim et al. [11] models user preferences as vectors to guide the enhancement process, yet the preparation of preference vectors is complicated. Furthermore, except for brightness information, color information is also learned in the vector, which leads to color distortion.

* Corresponding author at: School of Artificial Intelligence, Beijing University of Posts and Telecommunications, Beijing, China; Beijing Key Laboratory of Network System and Network Culture, Beijing University of Posts and Telecommunications, Beijing, China.
E-mail address: jiangzhuqing@bupt.edu.cn (Z. Jiang).
[1] The first two authors contribute equally to this work.

https://doi.org/10.1016/j.patcog.2022.108867

Fig. 1. Examples of multi-level enhancement employing our method. Two images in the middle are the two light levels enhanced results for low-light images, where the bottom-left corner is a brighter version compared to the top-right corner.

There is an inherent issue in personalized low-light image enhancement: human perception of brightness is non-linear, resulting in a misalignment between brightness values and subjective perception. To solve the misalignment, this paper proposes a deep learning algorithm for multi-level low-light image enhancement guided by arbitrary images serving as brightness references. Fig. 1 shows an example. Inspired by style transfer, we assume that an image consists of a content component and a luminance component in the latent space, which is later proved to be reasonable in our experiments. Specifically, content components refer to scene-invariant information during the enhancement, while luminance components represent brightness-specific information. Our method also delivers an additional advantage in alleviating color distortion.

Our method is similar but not identical to Retinex [12], which decomposes an image into two sub-images, namely reflectance and illumination. It enhances low-light images by adjusting the illumination and then recombine it with the corresponding reflectance. In contrast, our feature components are low-coupling, which allows a new image generated by concatenating two feature components from different images.

Our main contributions are summarized as follows:

(1) We propose a novel network for personalized low-light image enhancement, which interacts with users to flexibly conform to different needs and preferences. Specifically, the network takes references provided by users as targets and enhances low-light images to similar brightness.
(2) Our network decomposes images into low-coupling content components and luminance components in the latent space. Then the network combines the content features of low-light images and luminance features of arbitrary target images to perform multi-level low-light image enhancement.
(3) Our method alleviates color distortion. Furthermore, the multi-level enhancement effect is noticeable. Extensive experiments demonstrate its superiority over state-of-the-art methods on various datasets.

2. Related work

The proposed method is inspired by style transfer, which decouples an image into content and style. Similarly, our network decomposes an image into two feature components in the latent space for multi-level low-light enhancement. This section introduces the milestones in these two areas.

2.1. Low-light image enhancement

As the main task of this paper, we first introduce the development process of low-light image enhancement, including but not limited to works mentioned next.

2.1.1. One-to-one mapping

Conventional Methods. In the early work, histogram equalization and Retinex-based methods are the most widely adopted algorithms. As a precursor, histogram equalization and its variants concentrate on increasing image contrast. These methods adjust the histogram by spreading the most frequent intensity values to expand the limited dynamic range. However, such a globally balancing way generally leads to under- or over-enhancement. Another thinking, Retinex theory, treats an image as a pixel-wise product of two sub-images, called reflectance and illumination. The reflectance is assumed to be consistent under different lighting conditions, and the illumination is estimated to enhance under-lit images. Wang et al. [13] proposed a method for non-uniformly illumination images called NPE, which balances details and naturalness while improving contrast. Guo et al. [14] proposed LIME, a coarse-to-fine illumination estimation method. The illumination is first obtained by finding the maximum value of each pixel in RGB channels and then refined by imposing a structural prior. Li et al. [15] built a robust Retinex model that takes the inherent noise problem in low-light images into consideration. These methods are an improvement over earlier single-scale Retinex (SSR) and multi-scale Retinex (MSR), making enhanced images more natural and alleviating overexposure and slight noise. However, the color distortion intrinsic in low-light images remains unsolved. Also, complicated mapping of input and output is challenging to address.

Deep Learning-based Methods. In recent years, researchers pay more attention to deep learning and apply it to low-light image enhancement. Various deep models are proposed to learn the non-linear mapping of degraded low-light images to high-quality enhanced images. Deep learning-based methods are mainly classified as codec-based, Retinex-based and GAN-based methods. The codec-based method gradually reduces and restores the resolution of images to obtain semantically reliable contextual information. Lore et al. [6] proposed an autoencoder (LLNet) for simultaneously addressing low-light enhancement and denoising. Chen et al. [8] proposed an end-to-end training model to light up extremely dark images, which operates on raw images. Guo et al. [16] introduced a method to complete the enhancement mapping by generating an image-specific curve instead of directly outputting an image. Zamir et al. [9] designed a deep model to maintain spatial details and receive strong contextual information based on a set of multi-scale features. Yang et al. [17] designed a semi-supervised learning approach to recover and improve a linear band representation with both paired and unpaired data. Kar et al. [18] proposed a zero-shot model, which estimates Koschmieders model parameters related to the degradation. The Retinex-based method utilizes neural networks to decompose images and then manipulates the reflectance and the illumination. Wei et al. [7] proposed Retinex-Net, which combines conventional Retinex theory with deep networks. Shen et al. [19] constructed MSR-net that established a relationship between multi-scale Retinex and feedforward convolutional neural networks to learn an end-to-end mapping from dark to bright images. Wang et al. [20] designed a deep model to learn an image-to-illumination mapping, rather than an image-to-image mapping in most methods. Fan et al. [21] integrated the idea of Retinex decomposition and semantic segmen-

tation to use semantic layer information to enhance separate regions. Liu et al. [22] proposed a lightweight network, which builds fundamental architectures and automatically discovers the embedded prior architectures for different scenarios. Zhang et al. [23] introduced a self-supervised method to reduce noise and improve contrast to avoid the blur caused by pre-/post-denoising. The GAN-based method eliminates the limitation of paired image training in contrast with the above two methods. Deng et al. [24] introduced a weakly supervised model by applying binary labels on image aesthetic quality to adversarial learning. Jiang et al. [25] proposed an unsupervised generative adversarial network (EnlightenGAN) without low/normal-light image pairs for training. Yang et al. [26] perform a network trained with paired and unpaired data to enhance contrast and reduce noise simultaneously.

It is particularly pointed that although our method and the Retinex-based method both decompose the image into two components, the decomposition space is different. Retinex theory assumes that an image can be decomposed into the pixel-wise product of reflectance R and illumination I. We assume that a feature of an image can be decomposed into the concatenation of content c and luminance l. The former decomposition is in the image space, while the latter is in the feature space. The above is a unified difference between these two ways, and there are more differences when specifying a particular model. For example, Wang et al. [20] regards R in Retinex theory as a well-exposed image and estimates I to obtain result R. Since R is constant under different brightness conditions, Wang et al. [20] produces one result for each low-light image. In contrast, we re-concatenated c and various l containing different brightness information to produce multiple results for each low-light image.

These methods have extensively promoted the development of low-light enhancement technology. On the basis of using different theories to light up objects well, they also try to solve other common degradations, such as noise and blur, to obtain a more satisfactory quality and vision perception. They are either trained with paired images to learn brightness difference between image pairs or trained with unpaired images whose target domain images or labels are carefully selected. Unfortunately, such ways cause models to only generate results with a fixed brightness, and the mapping between low-light and enhanced images is one-to-one. Therefore, none of these methods support a selection of brightness levels during enhancement, whether guided by a given value or an image. However, the results are expected to be diverse, given application scenarios and aesthetic standards. Thus, subjectivity needs to be emphasized by introducing these factors as a guideline in the enhancement process. It is practical to realize a one-to-many mapping to address multiple but specific requirements.

2.1.2. One-to-many mapping

Gamma correction is an adaptable image enhancement method, which performs non-linear operations on pixel values. The individual operation on each pixel ignores the relationship between adjacent pixels. Kang et al. [27] proposed an enhancement system with a set of adjustable parameters. The parameters were obtained by enhancing 25 representative images. For a new input, the most similar representative image is searching through metric learning, and then the system adopts its corresponding parameters to enhance the input image. However, the similarity is biased since 25 representative images are incapable of covering all situations. Besides, the parameters only represent the aesthetic of the user manipulating the images. Recently, Zhang et al. [10] proposed a deep network with an interface for practically enhancing low-light images, where a strength ratio α is specified to manipulate the decomposed illumination. Nevertheless, α is not a useful descriptor since the strength ratio and the perceived change of light level are non-linear. Kim et al. [11] developed PieNet to enhance images adaptively introducing each user's preference vector. For a new user, the network asks him or her to select about 10∼20 preferred images and represents them as a vector in an embedding space. The color distortion appears in results as color information is introduced except for brightness information when modeling the preference vector.

2.2. Style transfer

Style transfer renders an input image in the style of a reference image while preserving the content, which is developed from non-photorealistic rendering (NPR) to neural style transfer (NST). NPR includes image synthesis from 3D models, the emulation of substrate and media, and user interaction. The seminal work of NST is that Gatys et al. [28] proposed an algorithm to reproduce a given painting style on a natural image, which was achieved by iteratively optimizing an input variable image to match the content of a photo and the style of a painting. The follow-up studies promoted slow optimization to fast transfer. The limitation afterward is each network trained for a fixed style. To address this problem, Chen et al. [29] learned more general style representations. So far, style transfer has been extended to domain transfer to deal with multiple tasks, such as portrait painting style transfer, visual attribute transfer [30] and video style transfer [31]. The key to these methods is to disentangle the content and style representations of images.

Inspired by style transfer, this paper proposes a multi-level low-light image enhancement method guided by arbitrary brightness reference. The proposed method and style transfer share a common principle that reduces the complexity of problems through decoupling. Our network decomposes an image into the content component and luminance component, which similarly corresponds to content and style in style transfer. Although both ways contain the content components, they refer to different concepts. Specifically, the *content* represents the spatial structure in style transfer while referring to all information except brightness in our method. Besides, unlike the style in style transfer which contains various information such as colors, textures and shapes, the luminance component only represents the light level in our task. Our paradigm is to input two images, one as an image to be enhanced to provide content information, and the other as a reference to indicate the light target level.

3. Methodology

The goal of low-light image enhancement is to learn a mapping from an image to a normal-light version. However, the NORMAL light level is within a range rather than a discrete value from both qualitative and quantitative perspectives. Thus, it is suggested the enhancement is a one-to-many mapping given application scenarios or users' aesthetic. To achieve multi-level low-light image enhancement, we make basic assumptions in Section 3.1. Then the network structure and loss function used to optimize the network are described in detail in Sections 3.2 and 3.3 respectively. Finally, we introduce details of prediction in Section 3.4.

3.1. Assumptions

Assumption 1. An image can be decomposed into two feature components in the latent space, namely the content component and the luminance component.

Let $\vec{x} = \{x_1, x_2, \ldots, x_n\}$ be a set of images with different light levels in the same scene. For each image x_i, f_i is its feature

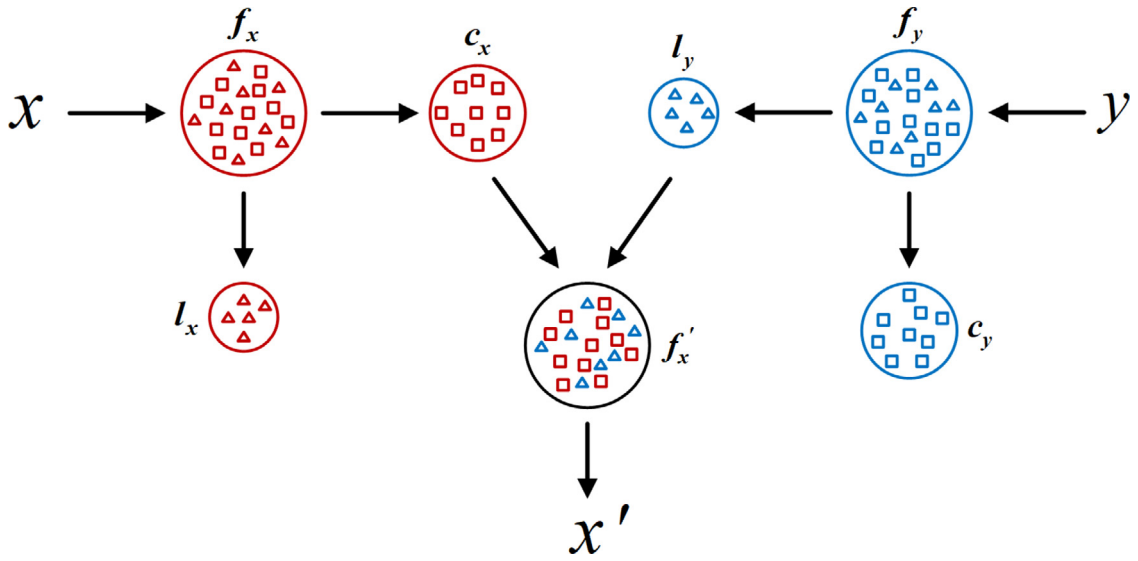

Fig. 2. A schematic diagram of assumption 2. The image pair (x, y) with irrelevant content is decomposed and concatenated in the latent space to generate an enhancement result of x.

vector in the latent space, which consists of a content component c and a luminance component l_i. In our assumption, c is invariant while l_i is specific for light levels i. In other words, a pair of images (x_i, x_j), where $i \neq j$, are encoded by an encoder E to generate feature vectors $f_i = E(x_i)$ and $f_j = E(x_j)$. In the latent space, f_i and f_j are decomposed into (c_i, l_i) and (c_j, l_j). Next, c_i and l_j are concatenated to form a new feature vector f_i', then $f_i' = f_j$. The reconstructed image of f_i' by a decoder G should be the same as x_j. In this way, multi-level mapping is performed by extracting luminance components from images with diverse light levels.

Assumption 2. Two feature components with fixed dimensions are low-coupling.

The above $\vec{x}$ is challenging to acquire in practice, so it is considered to use images that are content-irrelevant with the low-light images as a guideline. ***This paper executes the multi-level low-light image enhancement task guided by arbitrary images as brightness references regardless of scenes.*** Thus, the components are expected to be low-coupling to concatenate two images without involving information independent of brightness in the reference image. As shown in Fig. 2, let (x, y) be an image pair with different scenes, where y is a brightness reference image. The goal of the task is learning a mapping from x to a normal-light version x' which is as bright as y. Specifically, the feature vectors of x and y are decomposed into (c_x, l_x) and (c_y, l_y) respectively, and then c_x and l_y are concatenated to reconstruct an enhancement result of x. The result preserves original scene-invariant information of x and introduces target brightness from y. By taking different reference images as guidance, multi-level low-light image enhancement is achieved. The key to testifying assumptions is learning an encoder E and a decoder G using convolutional neural networks.

We first assume that an encoded feature vector containing all image information can be decomposed into several components containing sub-information. And then, we assume brightness information that is initially contained but randomly distributed in feature vectors can be present as a regular distribution. These two assumptions have no particular restrictions on images and no modification of information. Thus, they are applicable.

3.2. Architecture

Our model is designed to enhance a low-light image to corresponding normal-light versions. We present the network structure in Fig. 3. It consists of an encoder E, a feature concatenation module, and a decoder G, which form a U shape. The network takes two images as input, including a low-light image I_i and a reference image I_r. During training, I_r and I_i are identical in content, while in testing, I_r is an arbitrary image. Both inputs share the parameters of E.

As shown in Fig. 4, our network employs the down-sampling part of U-Net [32] as E, followed by a global average pooling operation, which respectively encodes I_i and I_r as feature vectors f_i and f_r. Correspondingly, G is formed by a fully connected layer and the up-sampling part of U-Net to reconstruct the feature vector. Details about the feature concatenation module are then provided, which is a crucial part of our network. In the encoder E and the decoder G, the size of other convolutional kernels is 3×3 except for the last convolutional layer. Each 3×3 convolution is followed by a rectified linear unit (ReLU), which can be expressed as:

$$f(x) = \max(0, x) \tag{1}$$

Feature concatenation module

Its function is to regroup components from two input feature vectors so that the output vector contains all desired information. Specifically, f_i and f_r are fed into the feature concatenation module, and their components, *i.e.* c_i and l_r, are concatenated to obtain a new feature f_i'. The concatenation of feature components can be expressed as follows:

$$f_i' = c_i \oplus l_r \tag{2}$$

Finally, the module produces a concatenation feature vector, which is then transformed into a feature map through a fully connected layer and dimension expansion operation. The feature map has the same resolution and channels as the corresponding feature map in the encoding stage.

The low-light image is enhanced by introducing l_r while retaining c_i. This way alleviates the problem of color distortion and accords with the essence of the task, that is, only light level changes.

As stated in Section 3.1, input feature vectors of the module are decomposable, and decomposed components are low-coupling.

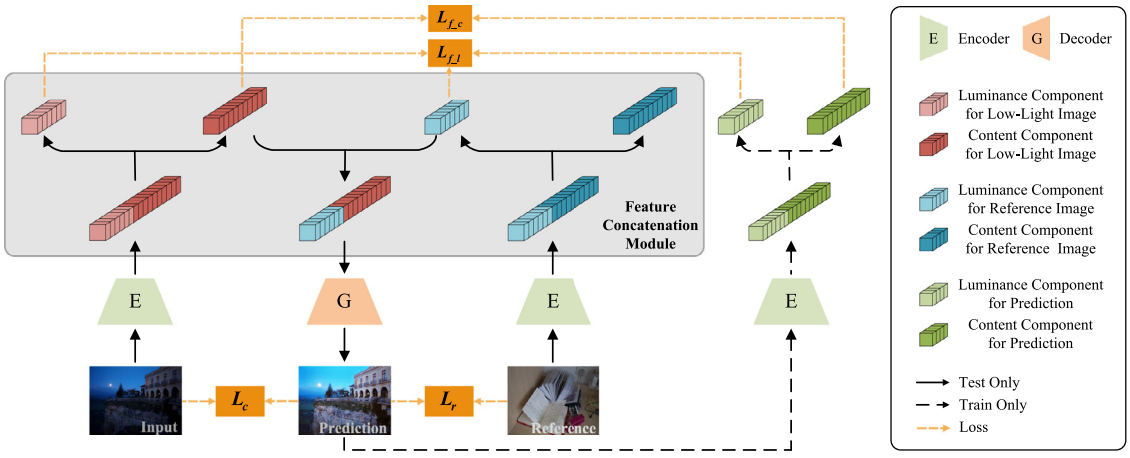

Fig. 3. An overview of the proposed framework. The two inputs (low-light and reference image) and the prediction (only available during training) share the encoder E parameters. During training, the inputs are content-identical image pairs, while the reference image can be arbitrary during testing. The network utilizes three losses for optimization, including reconstruction loss L_r, feature loss (L_{f_l} and L_{f_c}), and content consistency loss L_c.

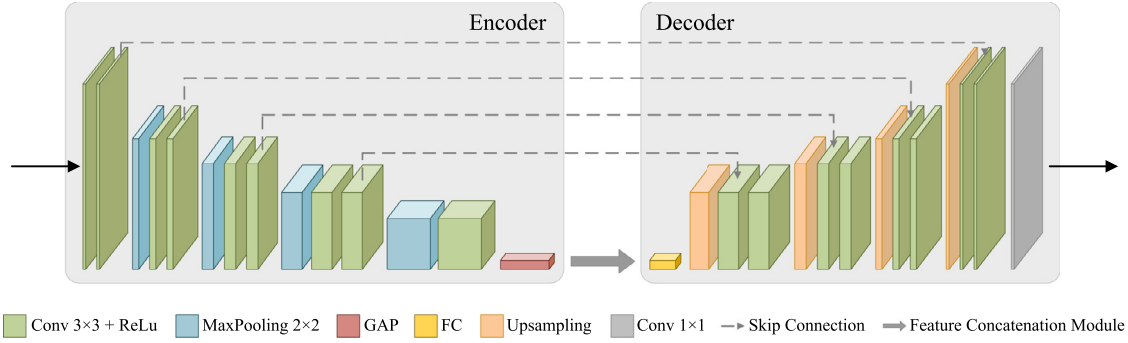

Fig. 4. The structure of the encoder E and decoder G. We employ the down-sampling part of U-Net followed by a global average pooling operation as the encoder E. The decoder G consists of a fully connected layer and U-Net up-sampling part.

Theoretically, for content-identical but brightness-distinct image pairs, the luminance components are wildly different, while the content components are approximately the same. And for content-irrelevant but brightness-identical images, the opposite is true. Therefore, the proposed method uses loss functions described in Section 3.3 to limit fixed dimensions of the vectors to include brightness information alone, and remaining dimensions include other information such as color, structure and details. These two kinds of information are non-overlapping.

3.3. Loss function

To perform the task, we propose several differentiable losses to restrict image-to-image and feature-to-feature processes. The following three losses are minimized to train our network.

Reconstruction loss In the image-to-image process, we compute the reconstruction loss. The L_1 error is used to measure the distance between the prediction and the ground truth. The reconstruction loss can be expressed as:

$$L_r = \| G(c_i, l_r) - I_r \|_1 \tag{3}$$

where I_r is the reference normal-light image, c_i is the content component decomposed by the low-light image I_i, and l_r is the luminance component decomposed by I_r. Pixels of all channels in the inputs of the network are normalized to [0,1].

The loss ensures that the network decomposes image pairs (I_i, I_r) with the same content into identical content components and different luminance components, which is achieved by reconstructing the feature vector composed of c_i and l_r into an image consistent with I_r.

Feature loss The feature loss is designed for feature-to-feature mapping. It is expected that feature components are reconstructed after passing through the decoder and encoder. To this end, we use content feature loss and luminance feature loss to constrain and learn the reconstruction and extraction processes of feature components. The feature loss is expressed as:

$$L_f = L_{f\_c} + L_{f\_l} \tag{4}$$

Here, L_{f_c} and L_{f_l} are the content feature loss and the luminance feature loss. Specifically, the content feature loss is defined as:

$$L_{f\_c} = \| c_p - c_i \|_2 \tag{5}$$

where c_i and c_p represent the content components of the low-light image and the prediction. $\| \cdot \|_2$ is the L_2 error. The loss encourages the content component to be consistent with the original after decoding and encoding, which is equivalent to ensuring that the content of the image remains unchanged after enhancement. Next, we refer to the definition of the triplet loss to define the luminance feature loss as:

$$L_{f\_l} = [\mathcal{D}(l_p, l_r) - \mathcal{D}(l_p, l_i) + \alpha]_+ \tag{6}$$

where l_i, l_r, and l_p respectively represent the luminance components of the low-light image, reference image, and the prediction.

$[\cdot]_+$ is a rectifier. The loss is the value in the rectifier when it is greater than 0; otherwise, the loss is 0. $\mathcal{D}(,)$ is the squared Euclidean distance between feature vectors. α is a margin. It averages the distances of the luminance components of 20 image pairs which are randomly selected from the dataset. α is calculated as:

$$\alpha = \sum_{i=1}^{20} \mathcal{D}(l_i^l, l_i^h) \tag{7}$$

where l_i^l and l_i^h represent the luminance components of the i-th image pair.

We choose triplet form rather than the L_2 metric used in the content feature loss. The reason is that l_p is expected to be similar to l_r and different from l_i on account of the specificity of the luminance component.

Content consistency loss Next, the content consistency loss is employed to restrict the enhanced image to be the same as the original low-light image except for the light level. Images are first mapped to the HSV color space. The optimization process penalizes the cosine distance of H and S channels between the prediction and the low-light image. The content consistency loss is expressed as:

$$L_c = L_{c\_H} + L_{c\_S} \tag{8}$$

Here, L_{c_H} and L_{c_S} respectively represent the cosine loss of H and S channels expressed as:

$$L_{c\_H} = 1 - \angle(H_p, H_i) \tag{9}$$

$$L_{c\_S} = 1 - \angle(S_p, S_i) \tag{10}$$

where H_i and H_p are the H channel of the low-light image and prediction, respectively. Similarly, S_i and S_p are the S channel. $\angle(,)$ is an operation to calculate cosine similarity, which is obtained by taking the dot product of two vectors' modulo. The closer this value is to 1, the more similar the two vectors are.

We support such color space mapping based on the following experiments. If the H and S channels of the low-light image are combined with the V channel of the content matching normal-light image, and then mapped back to the RGB space, the result is nearly the same as the normal-light image. It proves that the similarity of the H and S channels between the prediction and the low-light image can measure whether scene-invariant information changes after enhancement.

Cosine loss is adopted instead of the L_1 loss for the following reasons. First, the L_1 metric has been calculated in the RGB color space, which fails to figure whether the directions of pixel values are the same. It is also experimentally observed that the enhanced image color is closer to the ground truth when using the cosine loss than the L_1 loss.

Total loss The proposed network is optimized using the total loss:

$$L_{total} = L_r + \lambda L_f + L_c \tag{11}$$

where λ is a weight of the corresponding loss term. It is determined by trial-and-error. The weights are updated by back propagation to minimize the loss function. We use the Adam optimizer, which can be expressed as:

$$\theta_{t+1} = \theta_t - \frac{\eta}{\sqrt{v_t + \varepsilon}} * m_t \tag{12}$$

$$m_t = \frac{\beta_1 m_{t-1} + (1-\beta_1) g_t}{1 - \beta_1^t} \tag{13}$$

$$v_t = \frac{\beta_2 v_{t-1} + (1-\beta_2) g_t^2}{1 - \beta_2^t} \tag{14}$$

where η represents the learning rate. m_t is the first order momentum and v_t is the second order momentum. g_t represents the gradient. ε is a constant that prevents the denominator from being 0.

3.4. Details of prediction

A low-light image and a brightness reference are passed through E to get two feature vectors and then fed into the feature concatenation module. The module regroups c_i and l_r as the input of G. After decoding, we finally obtain an enhanced image of which the brightness is promoted to be similar to the reference.

4. Experiments

In this section, we begin with a training dataset brief and the training details. For the enhancement test, we first explore the dimension partitioning of the feature components to optimize the model performance. Then, the proposed method is compared with state-of-the-art methods through extensive qualitative and quantitative experiments. Besides, we conduct experiments related to the loss function to verify its effectiveness. Moreover, the ability of generating multi-level enhancement is demonstrated with arbitrary brightness references.

Dataset. LoL dataset [7] is involved in training. It consists of 500 image pairs, where each pair contains a low-light image and its corresponding normal-light image. The first 485 image pairs are for training, and the remaining are for testing.

MIT-Adobe 5K [33] dataset contains 5000 input images and 25,000 rendered images generated by five photographers (namely A/B/C/D/E). Thus, there are five sets of data, each of which is composed of 5000 low-light images and 5000 retouched images from one of the photographers. To be consistent with the previous method [9], we take the enhanced images of photographer C as ground truth. The first 4500 image pairs are used for training and the last 500 for testing.

Implementation Details. Our network is implemented with Tensorflow on an NVIDIA 2080Ti GPU. The model is optimized using Adam with a fixed learning rate of 1e−4. The batch size is set to 8. We train the model for 1000 epochs with a whole image as input. For data augmentation, a horizontal or vertical flip is randomly performed. Besides, a 100×100 image patch is stochastically located from each low-light image and is replaced with an image patch at the same position from the ground truth. The weight λ is set to 2, and the margin α is set to 0.08.

The network is trained in an end-to-end manner. During the training, a low-light image and its reference image are taken as input. After passing through our model, an enhanced image is generated, which is also fed into the encoder. The feature concatenation module produces feature components of the three images for the feature loss calculation.

4.1. Dimensions of feature components

According to Section 3.1, the low coupling of feature components means that scene-invariant and brightness-specific information is non-overlapping. Thus, our intention is to find the optimal dimension partitioning for the two kinds of information. Let the luminance component l be the first 32 to 480 dimensions with step size of 32 from 512-dimension feature vectors, and the content components are the remaining dimensions. The most appropriate setting is selected through comparisons of metrics.

Table 1 shows the four best dimension partitioning. We can see that PSNR is at least 0.27 dB higher than other cases when l is 96-dimensional. The same better performance under this setting is also reflected in SSIM. Based on the above results, our method fixes

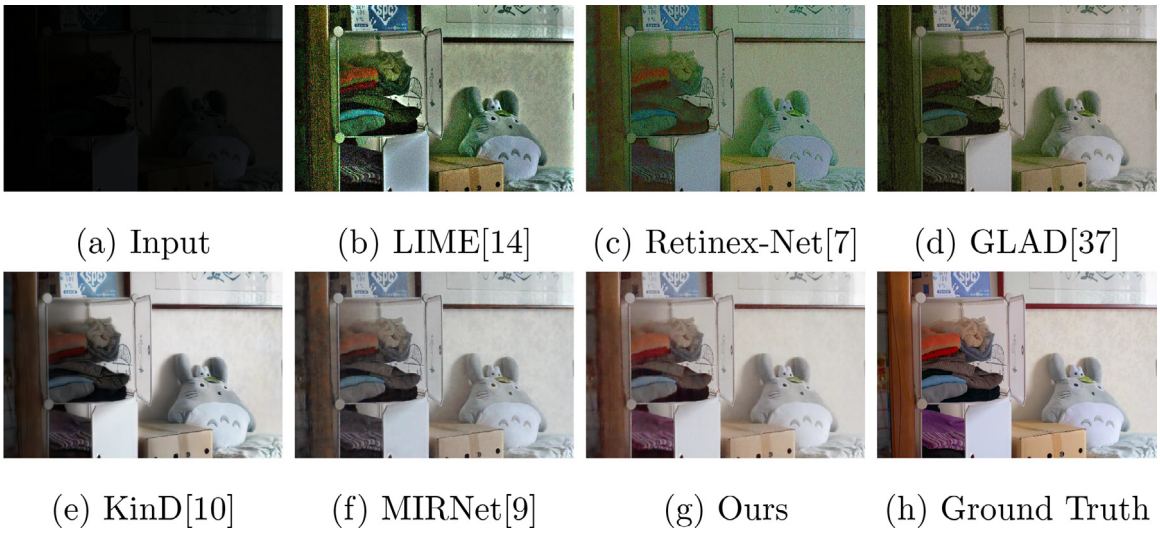

Fig. 5. Visual comparison with SOTA methods on the LoL dataset. Our result is more satisfactory than others in denoising, brightness enhancement and colors fidelity. On the one hand, the network retains the scene-invariant information of low-light images, leading to more natural colors and clearer details of results. On the other hand, the brightness-specific information introduced from reference images makes brightness levels of results more accurate.

Table 1
Qualitative comparison of feature components in different dimensions (32, 64, 96, and 128 dimensions) according to PSNR and SSIM. The best results are bolded.

Dimensions of l	PSNR	SSIM
32	27.42	0.857
64	27.47	0.859
96	**27.90**	**0.860**
128	27.63	0.858

the first 96 dimensions to be the luminance component l, and the remaining dimensions are the content component c.

Table 1 presents the principle of the method as well as results. For the principle, due to the limitation of loss functions on image-to-image and feature-to-feature processes and the special design of the feature concatenation module, the content component and the luminance component go into different layers of the feature vector. That is why we did not try other partitioning (such as mixed partitioning). For the results, the feasibility of decoupling presumptions is proved.

As explained in Section 3.2, content/brightness-identical images are supposed to have the same content/luminance components. However, in practice, light levels are not uniformly defined and are mostly determined by subjectivity. Thus, only the former case is verified by measuring similarities between feature components of image pairs. We calculate the mean square error (MSE) on the test set (15 image pairs) of the LOL dataset and take the mean value as the evaluation metric of the similarity. The mean MSE is 1.8×10^{-3} for luminance components and 8.7×10^{-5} for content components, proving the reasonability of the assumptions and the effect of the feature concatenation module.

4.2. Performance evaluation

Our method is tested on widely used datasets with or without ground truth. When the ground truth is present, the effectiveness of the proposed method is demonstrated by qualitative comparison on the LoL and MIT-Adobe 5K datasets with several state-of-the-art methods, such as KinD [10], MIRNet [9], RUAS [22] and PieNet [11]. Additionally, we conduct no-reference comparison on the LIME [14], DICM [34], NPE [13] and MEF [35] datasets.

Table 2
Low-light image enhancement evaluation on the LoL dataset. The best results are highlighted in red, and suboptimal results are highlighted in blue for PSNR and SSIM.

Method	PSNR	SSIM
LIME [14]	16.76	0.56
Retinex-Net [7]	16.77	0.56
GLAD [37]	19.72	0.70
KinD [10]	20.87	0.80
RUAS [22]	18.23	0.72
DRBN [17]	20.13	0.83
MIRNet [9]	24.14	0.83
Arora et al. [38]	23.01	0.89
Ours	27.90	0.86

Table 3
Low-light image enhancement evaluation on the MIT-Adobe 5K dataset. The best results are highlighted in red, and suboptimal results are highlighted in blue for PSNR and SSIM.

Method	PSNR	SSIM
HDRNet [39]	21.96	0.87
DPE [40]	22.15	0.85
DeepUPE [20]	23.04	0.89
RUAS [22]	20.83	0.85
MIRNet [9]	23.73	0.93
PieNet [11]	25.28	0.91
Ours	28.39	0.95

4.2.1. Full-reference image quality assessment

In quantitative comparison, PSNR (peak signal-to-noise ratio) and SSIM (structural similarity) [36] are adopted as evaluation metrics. For a fair comparison, we trained all the methods on the same data. Furthermore, methods involved in the comparison all employ the default training set and test set. Tables 2 and 3 show results of our method and the others on the LoL and MIT-Adobe 5K datasets. The best results are highlighted in red, and suboptimal results are highlighted in blue for each metric. As we can see from the tables, our network outperforms all the other methods. Notably, the proposed method achieves 3.76 dB better than MIR-

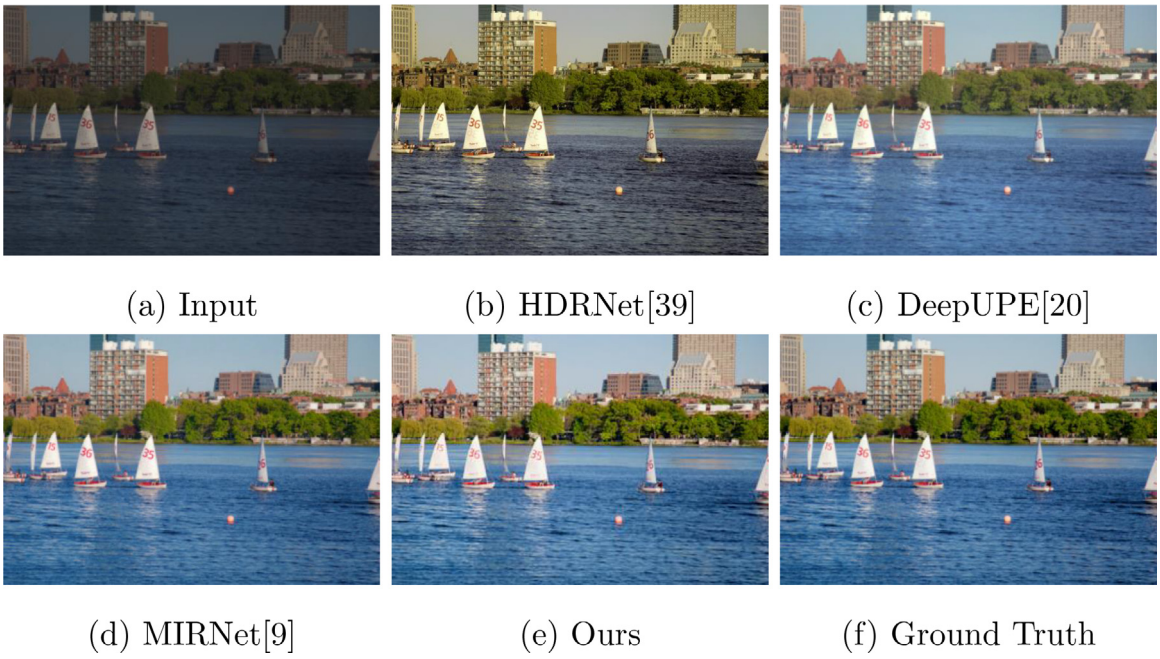

Fig. 6. Visual comparison with SOTA methods on the MIT-Adobe 5K dataset. Our result is more satisfactory than others in brightness enhancement and colors fidelity. On the one hand, the network retains the scene-invariant information of low-light images, leading to more natural colors and clearer details of results. On the other hand, the brightness-specific information introduced from reference images makes brightness levels of results more accurate.

Net on the LoL datasets and 3.11 dB better than PieNet on the MIT-Adobe 5K datasets, which are currently state-of-the-art. Figs. 5 and 6 visualize some examples. In Fig. 5, LIME, Retinex-Net and GLAD have insufficient brightness enhancement and more serious noise. KinDs texture is too smooth. KinD and MIRNet suffer from color distortion. In Fig. 6, the brightness of HDRNet is lacking. Color distortion appears in DeepUPE and MIRNet. In contrast, our method enhances dark regions and makes colors of the enhanced image closer to the ground truth.

The outperformance is attributed to two reasons. First, feature concatenation retains the scene-invariant information of the low-light image to the greatest extent, alleviating color distortion. Second, well-designed loss functions improve the performance of our network. Specifically, the content consistency loss constrains the scene-invariant information of low-light images and enhanced results to be identical in the image space, while the content feature loss does the same thing in the latent space. The luminance feature loss makes the brightness enhancement more accurate.

4.2.2. No-reference image quality assessment

We also compare different methods on four datasets without normal-light references, i.e., LIME, DICM, NPE, and MEF. We adopt NIQE (natural image quality evaluator) as a no-reference metric. A smaller value indicates a better result. Table 4 reports NIQE scores of the methods on the aforementioned datasets. The best results are highlighted in red, and suboptimal results are highlighted in blue. Our method has a higher NIQE of 0.03 than GLAD on the NPE dataset but a lower NIQE of at least 0.24 than others on the remaining three datasets. In the absence of ground truth, as can be seen from the results of different methods shown in Fig. 7, our method outputs more realistic and natural images. In contrast, LIME and Retinex-Net both introduce degradations, such as noise. Besides, LIME performs over-enhancement and unnatural colors, and Retinex-Net appears texture distortion. KinD fails to promote brightness to get accurate details. In a word, the proposed method achieves shaper contrast, richer colors and clearer details.

Table 4
The NIQE socres of different methods on four datasets (LIME, DICM, NPE, MEF). The best results are highlighted in red, and suboptimal results are highlighted in blue.

Method	LIME	DICM	NPE	MEF
LIME [14]	4.16	3.03	3.64	4.55
Retinex-Net [7]	5.95	4.71	4.29	5.63
KinD [10]	4.41	3.39	3.51	4.24
GLAD [37]	3.90	3.09	3.25	3.66
Ours	3.51	2.74	3.28	3.42

Table 5
Comparison of the luminance feature loss in different forms according to PSNR and SSIM. The best results are bolded.

Forms of L_{f_c}	PSNR	SSIM
L_2 loss	27.48	0.856
triplet loss	**27.90**	**0.860**

The principle behind the success of our method is that the network decouples the content component and the luminance component. The retention of the content component enables clearer details and more natural colors. Besides, the network allows to introduce different luminance components during enhancement. Our design of multiple trials provides accessibility to more satisfying results.

4.3. Effectiveness of loss function

We compare results when applying different features loss forms and select the better one as the final application. Furthermore, we also conduct an ablation study to demonstrate the effectiveness of each loss component. Quantitative results are reported in Tables 5 and 6.

Fig. 7. Visual comparison without ground truth. The first column images are respectively from the MEF, NPE, DICM and LIME datasets and under different lighting conditions. Our result is more satisfactory than others in denoising, brightness enhancement and colors fidelity. On the one hand, the network retains the scene-invariant information of low-light images, leading to more natural colors and clearer details of results. On the other hand, the brightness-specific information introduced from reference images makes brightness more consistent with subjective perception.

Table 6
Quantitative results of ablation study on the effectiveness of each loss component. The best result is bolded.

L_r	✓		✓	✓
L_f	✓	✓		✓
L_c		✓	✓	✓
PSNR	27.58	12.65	27.53	**27.90**

4.3.1. Forms of the feature loss

L_2 loss is widely used when constraining the similarity of two feature vectors. Nevertheless, in our method, we expect the similarity of the luminance components of results and references, and the dissimilarity between that of results and low-light images. Thus, triplet loss is a better choice. We compare the effects that the two forms bring on the network quantitatively in Table 5. The triplet loss achieves higher PSNR and SSIM, which means better enhancement.

4.3.2. Ablation study

In Table 6, we present comparisons when the model is trained using loss functions with different components. We treat the model of total loss components as the baseline. The result without reconstruction loss L_r has degradation of 15.25 dB in PSNR, demonstrating the importance of L_r in restoring images. It is also found that discarding feature loss L_f or content consistency loss L_c will degrade at least 0.32 dB in PSNR.

The above numerical results prove that each loss component is effective. On the other hand, L_c differs from L_r by constraining content consistency from H and S channels. Therefore, reconstructing image from a switched space improves the effect.

4.4. Different levels of enhancement

We show the results of multi-level mapping in Fig. 8. Our network generates multiple enhancement versions for the same low-light image, guided by various reference images. More importantly, the enhanced versions are basically the same in details, structures and colors. Furthermore, when different low-light images refer the same image, results have approximate brightness.

Most existing methods trained on paired datasets simply generate one fixed brightness result for a low-light image, which is a one-to-one mapping and means the lack of diversity. In contrast, our method achieves multi-level enhancement.

5. Conclusion

This paper focuses on the subjectivity of image enhancement and introduces brightness reference to help users get visually pleasing images. We propose a deep neural network for multi-level low-light image enhancement guided by arbitrary references. It can flexibly promote the brightness of low-light images to be similar to reference images without limitation of the scene. In addition, our method alleviates the problem of color distortion. Our network decomposes an image into two low-coupling feature components in the latent space. Then the content and luminance components of two images are concatenated to generate a new image. In this way, we can obtain multiple normal-light versions of one low-light image by selecting different reference images. Extensive experiments demonstrate the superiority of our method when compared with existing state-of-the-art methods. However, due to the high fidelity preservation of content information, denoising is still an open issue, which is part of our future work.

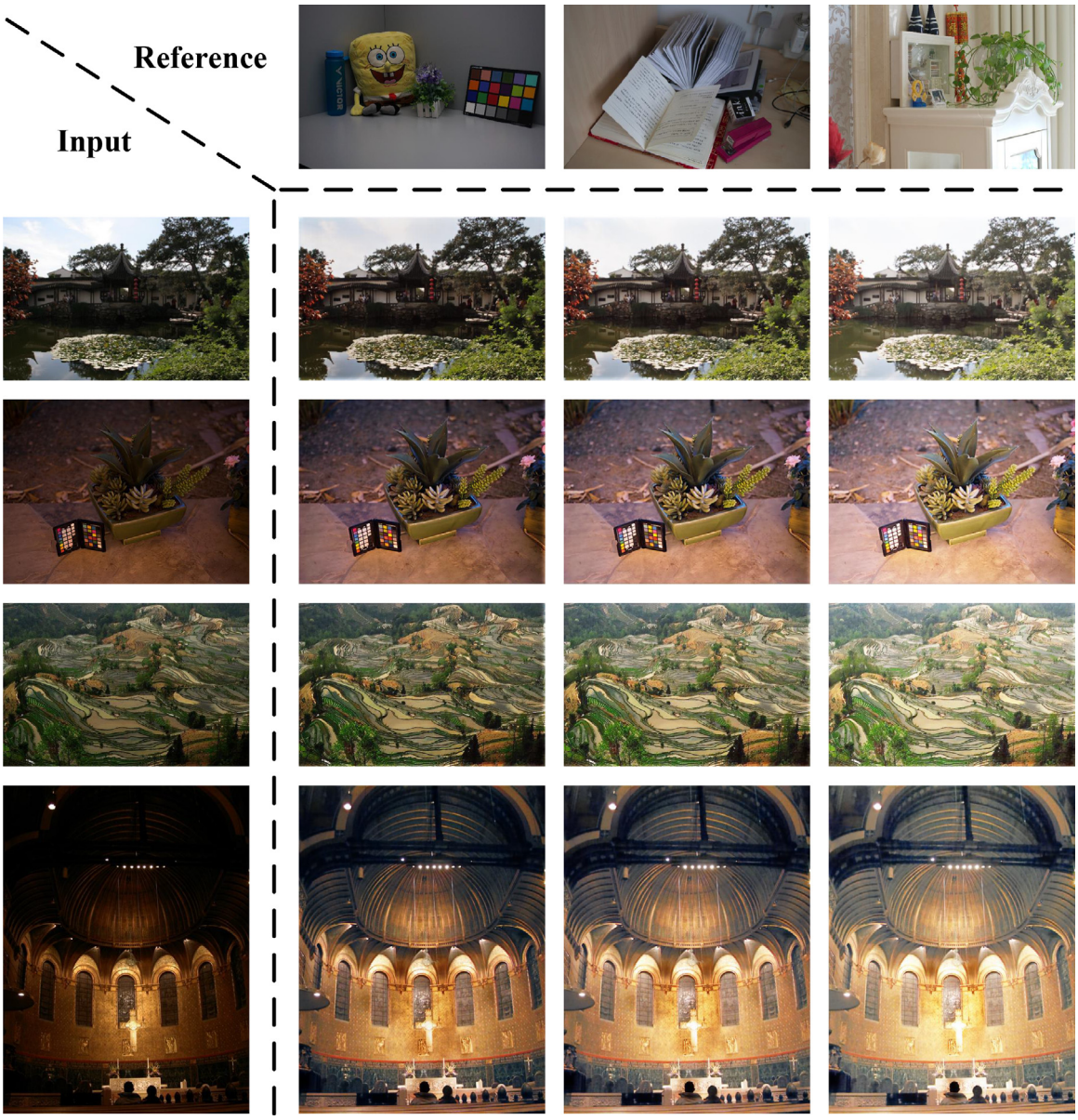

Fig. 8. Multi-level enhancement results. The ability of our network to generate multiple enhancement versions for the same low-light image is demonstrated. Since the network decouples the luminance component, different brightness information can be introduced into results to achieve multi-level enhancement. In addition, the retention of the scene-invariant information in low-light images enables multiple enhanced versions of an image to have the same details, structures and colors.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

Funding: This work was supported by the Ministry of Science and Technology of the People's Republic of China [grant number 2019YFC1511404]; and the National Natural Science Foundation of China [grant number 62002026].

Appendix A. Visualization of the feature component

We visualize the luminance and content features to verify the performance of the proposed network to decompose images in the latent space. However, re-concatenation in the feature concatenation module challenges visualization. Direct visualization before re-concatenation provides no semantic information since the feature components are low-dimensional embedding in the latent space. The visualized after re-concatenation is an enhanced image, and luminance and content cannot be analyzed separately. Thus, we conduct a detouring experiment leading to indirect observation. A white image is utilized to extract the content components and then combined with the illuminance components of natural images. Moreover, according to our assumption, an image with no content and gray shades affected by brightness levels will be obtained. In the experiment, we extract the content from the white image (a) and the luminance from images (b) and (c) in Fig. 9 to verify that no additional content information appears in the luminance component. The concatenated images are shown in (d) and (e). These two results only contain different brightness, which supports our belief that no content-related information be extracted from (b) and (c).

(a) White (b) Low-light (c) Normal-light (d) Result1 (e) Result2

Fig. 9. Visualization of the feature component.

References

[1] F. Yuan, Y. Zhou, X. Xia, X. Qian, J. Huang, A confidence prior for image dehazing, Pattern Recognit. 119 (2021) 108076.
[2] P.W. Hsieh, P.C. Shao, Blind image deblurring based on the sparsity of patch minimum information, Pattern Recognit. 109 (2021) 107597.
[3] X. Gao, Y. Wang, J. Cheng, M. Xu, M. Wang, Meta-learning based relation and representation learning networks for single-image deraining, Pattern Recognit. 120 (2021) 108124.
[4] Y. Zeng, Y. Gong, J. Zhang, Feature learning and patch matching for diverse image inpainting, Pattern Recognit. 119 (2021) 108036.
[5] W. He, Y. Chen, N. Yokoya, C. Li, Q. Zhao, Hyperspectral super-resolution via coupled tensor ring factorization, Pattern Recognit. 122 (2022) 108280.
[6] K.G. Lore, A. Akintayo, S. Sarkar, LLNet: a deep autoencoder approach to natural low-light image enhancement, Pattern Recognit. 61 (2017) 650–662.
[7] C. Wei, W. Wang, W. Yang, J. Liu, Deep retinex decomposition for low-light enhancement, arXiv preprint arXiv:1808.04560 (2018).
[8] C. Chen, Q. Chen, J. Xu, V. Koltun, Learning to see in the dark, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, 2018, pp. 3291–3300.
[9] S.W. Zamir, A. Arora, S. Khan, M. Hayat, F.S. Khan, M.H. Yang, L. Shao, Learning enriched features for real image restoration and enhancement, in: Proceedings of European Conference on Computer Vision (ECCV), Springer, 2020, pp. 492–511.
[10] Y. Zhang, J. Zhang, X. Guo, Kindling the darkness: a practical low-light image enhancer, in: Proceedings of the ACM International Conference on Multimedia (ACM MM), 2019, pp. 1632–1640.
[11] H.U. Kim, Y.J. Koh, C.S. Kim, PieNet: personalized image enhancement network, in: Proceedings of European Conference on Computer Vision (ECCV), Springer, 2020, pp. 374–390.
[12] E.H. Land, The retinex theory of color vision, Sci. Am. 237 (6) (1977) 108–129.
[13] S. Wang, J. Zheng, H.M. Hu, B. Li, Naturalness preserved enhancement algorithm for non-uniform illumination images, IEEE Trans. Image Process. 22 (9) (2013) 3538–3548.
[14] X. Guo, Y. Li, H. Ling, Lime: low-light image enhancement via illumination map estimation, IEEE Trans. Image Process. 26 (2) (2016) 982–993.
[15] M. Li, J. Liu, W. Yang, X. Sun, Z. Guo, Structure-revealing low-light image enhancement via robust retinex model, IEEE Trans. Image Process. 27 (6) (2018) 2828–2841.
[16] C. Guo, C. Li, J. Guo, C.C. Loy, J. Hou, S. Kwong, R. Cong, Zero-reference deep curve estimation for low-light image enhancement, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, 2020, pp. 1780–1789.
[17] W. Yang, S. Wang, Y. Fang, Y. Wang, J. Liu, From fidelity to perceptual quality: asemi-supervised approach for low-light image enhancement, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, 2020, pp. 3063–3072.
[18] A. Kar, S.K. Dhara, D. Sen, P.K. Biswas, Zero-shot single image restoration through controlled perturbation of koschmieder's model, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, 2021, pp. 16205–16215.
[19] L. Shen, Z. Yue, F. Feng, Q. Chen, S. Liu, J. Ma, MSR-Net: low-light image enhancement using deep convolutional network, arXiv preprint arXiv:1711.02488 (2017).
[20] R. Wang, Q. Zhang, C.W. Fu, X. Shen, W.S. Zheng, J. Jia, Underexposed photo enhancement using deep illumination estimation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, 2019, pp. 6849–6857.
[21] M. Fan, W. Wang, W. Yang, J. Liu, Integrating semantic segmentation and retinex model for low-light image enhancement, in: Proceedings of the ACM International Conference on Multimedia (ACM MM), 2020, pp. 2317–2325.
[22] R. Liu, L. Ma, J. Zhang, X. Fan, Z. Luo, 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 (CVPR), IEEE, 2021, pp. 10561–10570.
[23] Y. Zhang, X. Di, B. Zhang, Q. Li, S. Yan, C. Wang, Self-supervised low light image enhancement and denoising, arXiv preprint arXiv:2103.00832 (2021).
[24] Y. Deng, C.C. Loy, X. Tang, Aesthetic-driven image enhancement by adversarial learning, in: Proceedings of the ACM International Conference on Multimedia (ACM MM), 2018, pp. 870–878.
[25] Y. Jiang, X. Gong, D. Liu, Y. Cheng, C. Fang, X. Shen, J. Yang, P. Zhou, Z. Wang, EnlightenGAN: deep light enhancement without paired supervision, IEEE Trans. Image Process. 30 (2021) 2340–2349.
[26] Q. Yang, Y. Wu, D. Cao, M. Luo, T. Wei, A lowlight image enhancement method learning from both paired and unpaired data by adversarial training, Neurocomputing 433 (2021) 83–95.
[27] S.B. Kang, A. Kapoor, D. Lischinski, Personalization of image enhancement, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, 2010, pp. 1799–1806.
[28] L.A. Gatys, A.S. Ecker, M. Bethge, A neural algorithm of artistic style, arXiv preprint arXiv:1508.06576 (2015).
[29] D. Chen, L. Yuan, J. Liao, N. Yu, G. Hua, StyleBank: an explicit representation for neural image style transfer, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, 2017, pp. 1897–1906.
[30] D. Kotovenko, A. Sanakoyeu, P. Ma, S. Lang, B. Ommer, A content transformation block for image style transfer, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, 2019, pp. 10032–10041.
[31] M. Ruder, A. Dosovitskiy, T. Brox, Artistic style transfer for videos and spherical images, Int. J. Comput. Vis. 126 (11) (2018) 1199–1219.
[32] O. Ronneberger, P. Fischer, T. Brox, U-Net: convolutional networks for biomedical image segmentation, in: Proceedings of International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI), Springer, 2015, pp. 234–241.
[33] V. Bychkovsky, S. Paris, E. Chan, F. Durand, Learning photographic global tonal adjustment with a database of input/output image pairs, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, 2011, pp. 97–104.
[34] C. Lee, C. Lee, C.S. Kim, Contrast enhancement based on layered difference representation, in: Proceedings of IEEE International Conference on Image Processing (ICIP), IEEE, 2012, pp. 965–968.
[35] K. Ma, K. Zeng, Z. Wang, Perceptual quality assessment for multi-exposure image fusion, IEEE Trans. Image Process. 24 (11) (2015) 3345–3356.
[36] Z. Wang, A.C. Bovik, H.R. Sheikh, E.P. Simoncelli, Image quality assessment: from error visibility to structural similarity, IEEE Trans. Image Process. 13 (4) (2004) 600–612.
[37] W. Wang, C. Wei, W. Yang, J. Liu, GLADNet: low-light enhancement network with global awareness, in: Proceedings of IEEE International Conference on Automatic Face and Gesture Recognition, IEEE, 2018, pp. 751–755.
[38] A. Arora, M. Haris, S.W. Zamir, M. Hayat, F.S. Khan, L. Shao, M.H. Yang, Low light image enhancement via global and local context modeling, arXiv preprint arXiv:2101.00850 (2021).
[39] M. Gharbi, J. Chen, J.T. Barron, S.W. Hasinoff, F. Durand, Deep bilateral learning for real-time image enhancement, ACM Trans. Graph. 36 (4) (2017) 1–12.
[40] D.J. Jobson, Z.u. Rahman, G.A. Woodell, A multiscale retinex for bridging the gap between color images and the human observation of scenes, IEEE Trans. Image Process. 6 (7) (1997) 965–976.

Yanan Wang is a graduate student with the School of Artificial Intelligence, Beijing University of Posts and Telecommunications (BUPT), Beijing, China. She received the B.S. degree in communication engineering system from Beijing Jiaotong University (BJTU) in 2019. Her research interests include image processing and low-light image enhancement.

Zhuqing Jiang is a Lecturer with the School of Artificial Intelligence, Beijing University of Posts and Telecommunications (BUPT), Beijing, China. He received his Ph.D. degree in communication and information system from BUPT in 2014. His research interests include satellite networking and indoor localization.

Chang Liu received the B.S. degree from the Beijing University of Posts and Telecommunications Beijing, China, in 2019, where he is currently pursuing the master's degree. His research interests include image processing and image super-resolution.

Kai Li is a graduate student with the School of Artificial Intelligence, Beijing University of Posts and Telecommunications (BUPT), Beijing, China. He received his B.S. degree in the School of Information Science and Engineering from Shandong Normal University in 2019. His research interests include image processing and low-light image enhancement.

Aidong Men is a professor with the School of Artificial Intelligence, Beijing University of Posts and Telecommunications (BUPT), China. He received his Ph.D. degree in communication and information system from BUPT in 1994. His research interests include the transmission and processing of multimedia.

Haiying Wang is a female associate professor in the School of Information and Communication Engineering, Beijing University of Posts and Telecommunications (BUPT), China. She received the Ph.D. degree in communication and information system from BUPT in 2009. Her current interests include image and video processing, computer vision.

Xiaobo Chen is a senior engineer with the Supervision Center of National Radio and Television Administration (NRTA). He received his Ph.D. degree in communication and information system from BUPT in 2011. His research interests include datacenter architecture and video processing.