Structure-Texture Aware Network for Low-Light Image Enhancement

Kai Xu, Huaian Chen, Chunmei Xu, Yi Jin, *Member, IEEE*, and Changan Zhu

***Abstract*— Global structure and local detailed texture have different effects on image enhancement tasks. However, most existing works treated these two components in the same way, without fully considering the characteristics of the global structure and local detailed texture. In this work, we propose a structure-texture aware network (STANet) that successfully exploits structure and texture features of low-light images to improve perceptual quality. To construct STANet, a fine-scale contour map guided filter is introduced to decompose the image into a structure component and a texture component. Then, structure-attention and texture-attention subnetworks are designed to fully exploit the characteristics of these two components. Finally, a fusion subnetwork with attention mechanisms is utilized to explore the internal correlations among the global and local features. Furthermore, to optimize the proposed STANet model, we propose a hybrid loss function; specifically, a color loss function is introduced to alleviate color distortion in the enhanced image. Extensive experiments demonstrate that the proposed method improves the visual quality of images; moreover, STANet outperforms most other state-of-the-art approaches.**

***Index Terms*— Low-light image enhancement, structure-texture aware, guided filter, hybrid loss.**

I. Introduction

ACQUIRING natural scene images with good contrast, rich details, and vivid color is an essential goal of digital photography. However, captured images are easily underexposed in low-light environments, leading to decreased performance in some high-level vision tasks, such as object detection, semantic segmentation, and image classification [1]. To address this problem, numerous works have been proposed to enhance image quality. Early works [2]–[6] focused on contrast and saturation enhancement by adjusting the brightness of the whole low-light image. However, these methods only considered global features and thus lost considerable local details. Another line of research [7], [8] has focused on improving image quality based on the Retinex theory. These methods decompose the input image into an illumination map and a reflection map. As a result, the reflection map is taken as the final enhanced image. Compared with early works, Retinex-based methods achieve considerable improvements in image quality. However, these methods may amplify noise and cause over-enhancement.

Recently, benefiting from the superiority of convolutional neural networks (CNNs), many works have attempted to apply CNNs in image enhancement tasks. For example, Ignatov *et al.* [9] proposed an end-to-end residual network to learn a map between photos from mobile devices and DSLR cameras. Huang *et al.* [10] adopted an encode-decode architecture to learn the global features of input images and achieved improved enhancement results. Li *et al.* [11] designed a progressive-recursive network that uses a recursive unit to repeatedly unfold the input image for feature extraction, which produces pleasing results for low-light conditions. However, the image enhancement task is a position-sensitive procedure in which pixel-to-pixel correspondence from the input image to the output image is needed. Therefore, this task is very sensitive to texture details, and the network should not discard the details when extracting global information. The aforementioned methods adopt well-designed neural network architectures to exploit these components identically, which inevitably neglects the detailed texture in an enhanced image and thus results in a decrease in visual quality.

To fully consider the global structure and local detailed texture, we present a structure-texture aware network that adopts attention subnetworks to separately consider the structure and texture information. The proposed network consists of a structure-texture decomposition module and a structure-texture aware module. The structure-texture decomposition module utilizes a map-guided filter based on a fine contour map to decompose a degraded image into global structure maps and local texture maps. Global structure maps preserve large-scale structural features with shaper edges, whereas local texture maps contain fine details. The structure-texture aware module includes a structure-attention subnetwork, a texture-attention subnetwork, and a fusion subnetwork. The structure-attention subnetwork uses the structure maps as input and adopts an encoder-decoder architecture to extract the global features.

Manuscript received 19 October 2021; revised 20 December 2021; accepted 1 January 2022. Date of publication 7 January 2022; date of current version 4 August 2022. This work was supported in part by the National Natural Science Foundation of China under Grant 61727809, in part by the Special Fund for Key Program of Science and Technology of Anhui Province under Grant 201903c08020002, and in part by the National Key Research and Development Program of China under Grant 2019YFC0117800. This article was recommended by Associate Editor Z. Chen. *(Corresponding author: Yi Jin.)*

Kai Xu, Huaian Chen, and Chunmei Xu are with the School of Engineering Science, University of Science and Technology of China, Hefei, Anhui 230022, China (e-mail: xukaikai@mail.ustc.edu.cn; anchen@mail.ustc.edu.cn; xchm@mail.ustc.edu.cn).

Yi Jin and Changan Zhu are with the School of Engineering Science and the School of Data Science, University of Science and Technology of China, Hefei, Anhui 230022, China (e-mail: jinyi08@ustc.edu.cn; changan@ustc.edu.cn).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TCSVT.2022.3141578.

Digital Object Identifier 10.1109/TCSVT.2022.3141578

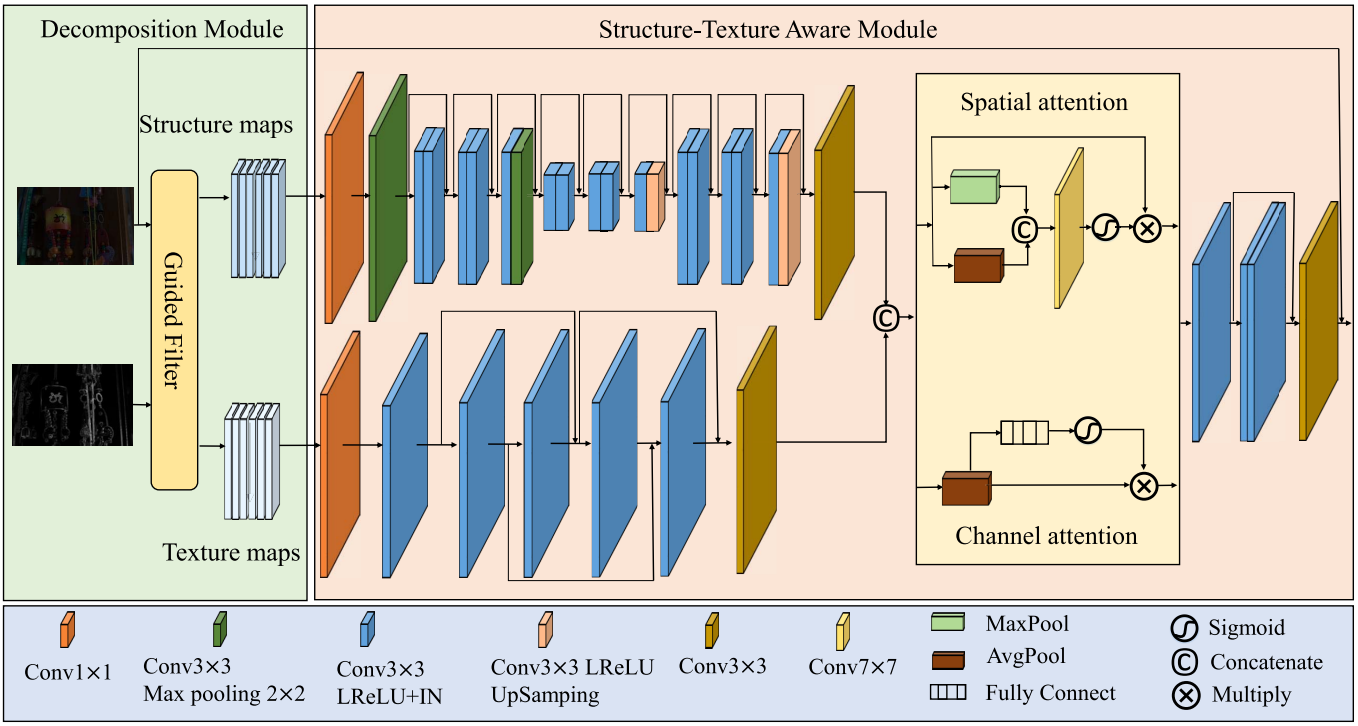


Fig. 1. The architecture of our proposed STANet model. The structure-texture decomposition module utilizes the contour map-guided filter to decompose the image into structure and texture maps. Then, two attention subnetworks are used to separately consider the structure maps and texture maps. Finally, a fusion network is adopted to generate an enhanced image by integrating the global structure information and local detailed information. Conv $k \times k$ denotes a convolutional layer with a kernel size of $k \times k$.

In contrast, the texture-attention subnetwork uses the texture maps and cascading residual blocks to focus on local features and recover details in an enhanced image. The fusion subnetwork utilizes spatial and channel attention mechanisms to integrate global and local features and to explore the internal correlations among the features. Thus, the proposed method successfully considers the structure and texture information, which allows STANet to retain abundant detailed information when extracting global structure information. Moreover, we propose a hybrid loss function to optimize STANet, where a color loss function is utilized to alleviate the color distortion in an enhanced image. In this way, STANet improves visual quality.

1) We propose a structure-texture aware network that successfully considers the global structure features and local texture features of a low-light image.
2) We design a low-light image decomposition method that uses a finer contour map as the guided map to decompose a low-light image into global structure maps that retain large-scale sructural features and local texture maps that contain fine details.
3) We introduce a hybrid loss function to optimize the proposed STANet model. In particular, a color loss function is designed to alleviate the color distortion in the enhanced image.

The remainder of this work is organized as follows. In Section II, the previous works related to image enhancement are reviewed. Then, in Section III, we introduce the principle of the proposed method. In Section IV, we evaluate the performance of STANet through extensive comparison experiments and present the analyses. Finally, Section V concludes the paper.

II. Related Work

A. Image-Decomposition Methods

Image decomposition is a challenging task and has been extensively studied in several decades [12]. Traditional decomposition methods can be divided into two branches: TV-based methods and image filtering-based methods. TV-based methods always use different functional norms to differentiate the structural and textural components of images [13], such as TV-L2 [14], TV-G [15], TV-L1 [16], TV-Hilbert [17], etc. As a typical example, method in [18] proposed a relative total variation (RTV) method, which decomposes images based on the observation that the aggregation result of the signed gradient values in a local window usually has larger absolute values for edges than for textures. The TV-based methods can effectively separate the structural and textural components. However, most of these methods encourage strong edges and suppress weak ones, while the weak ones may be also related to structures. Thus, to address this problem, Kim *et al.* [19] introduced a model by learning deep variational priors for structure maps, which successfully differentiates high-amplitude details from structure edges, and avoids halo artifacts.

Another kind of method is based on an image filter. The well-known ones include Gaussian filter, nonlocal filter, bilateral filter, and weighted least square (WLS) filter, etc. These filters contribute to achieving smooth details and

preserving salient edges. With further study on these filters, Liu *et al.* [20] used contour information of the filtered image to design a spatially adaptive gradient sparsity regularization to guide image reconstruction, which helps preserve edges and suppress artifacts. Zhu *et al.* [21] considered the scale information of the structure and texture, and proposed a filter that uses standard Laplacian pyramids to characterize edges with a simple threshold on pixel values, which can distinguish large-scale edges from small-scale details. In [22], Yin *et al.* proposed a side window filtering (SWF) technique that aligns the side or corner of a window with the pixel being processed, which can effectively solve the edge blur problem caused by the window on the edge. To overcome complex situations, in [23], Xu and Wang presented a texture filter that adjusts the window size for a pixel by checking similar pixels in its neighborhood. Thus, the window sizes can be adjusted by the distances from pixels to structures.

Recently, several methods have devoted to apply CNN models to image decomposition tasks [24]. Lu *et al.* [25] generated a large dataset by blending natural textures with clean structure-only images. Furthermore, they built a texture prediction network (TPN) that can predict the location and magnitude of textures. However, this method adopts synthetic samples for training, and cannot achieve competitive decomposition results for real images. To solve this problem, Fan *et al.* [26] proposed an unsupervised learning method that employs a residual network to optimize a regularized loss function on external samples. In [27], Gao *et al.* proposed a semi-supervised method that distinguishes the structural content from non-structural textures at the sematic level. In another way, Zhou *et al.* [28] proposed a structure-texture decomposition method (STDN) that considers the anisotropy of local gradients and the repeatability degree of signal patterns in a neighboring region, which helps decompose finer textures.

The aforementioned decomposition methods have achieved impressive performance for decomposing structures and textures. However, most of them are based on the priors presented in normal-light images. These priors may be inapplicable in low-light images, which results in failed decomposition for structures and textures. In this work, we propose a decomposition method based on a guided filter that uses a contour map to guide the decomposition of global structure maps and local texture maps. To extract contour map for the low-light images, we use an exponent to the local derivatives to obtain finer contour map, which helps the proposed method to effectively decompose the structures and textures of low-light images.

B. Image-Enhancement Methods

1) Traditional Methods: Histogram equalization (HE) is a typical and efficient image enhancement approach. HE-based methods [2], [4], [6] map the value of an input image into [0, 1] and attempt to redistribute the luminous intensity throughout the corresponding histogram. However, these methods only focus on image contrast enhancement, and such simple pixel statistics and redistribution process may result in the distortion of the enhanced images. Another popular traditional approach is based on a camera response model, which considers the entire image processing pipeline of the camera. These methods [29]–[31] estimate the exposure ratio map and then adjusted each pixel to obtain the enhanced high-quality image, however, they minimally consider the real illumination factor, which tends to lead to the visual vulnerability of the enhanced image and inconsistency with real scenes. Retinex theory [32] considers the intrinsic properties of an image and assumes that an image can be decomposed into an illumination map that represents the variation in light and a reflection map that is approximate to the enhanced image. Typically, single-scale Retinex (SSR) and multi-scale Retinex (MSR) methods [33] proposed generating an enhanced image by using Retinex theory to decompose a low-light image and then directly remove the illumination map. As early attempts to apply Retinex theory in image enhancement, these two methods achieved great success in quality improvement. However, directly removing the illumination map will also amplify the noise, which makes the enhanced image look unnatural. To address this problem, Fu *et al.* [8] presented a novel method that sequentially optimizes the reflection map after removing the illumination map, which enriches the details of the enhanced image and thus yields improved visual quality. Yu and Zhu [34] constructed a physical lighting model to describe the degradation of poor illumination images and then recovered a low-light image by solving the model. Chen *et al.* [35] proposed a contrast enhancement method that uses an entropy-preserving mapping prior to develop a linear model and generates a mapping curve to enhance the image contrast. Li *et al.* [36] introduced a noise map to enhance low-light images, which helped suppress noise. Although the aforementioned algorithms dramatically improve the enhancement quality, they perform poorly in cases with heavy noise and color distortion.

2) CNN-Based Methods: CNNs have been widely used in image restoration tasks such as image denoising [37]–[39], image dehazing [40], and image super-resolution [41]. Recently, several works [9], [42]–[50] have successfully applied CNN models in image enhancement. Shen *et al.* [43] developed an end-to-end network to learn a map between low-light and bright images. However, this method cannot overcome the degradation of color distortion. Chen *et al.* [44] proposed a pipeline for low-light image enhancement based on a fully convolutional network that can jointly deal with heavy noise and color distortion. However, this approach can only deal with an image in RAW format, which greatly limits its applicability. The method in [9] introduced a dataset in RGB format and presented an end-to-end deep learning approach for image enhancement; after applying this method, the quality of an image taken with a low-end camera was similar to that of an image taken by a DSLR. Furthermore, to achieve real-time processing, Liu and Jung [46] proposed a network that utilized multiple connected residual networks to improve the enhancement results and reduce the computational complexity. Ren *et al.* [47] presented a hybrid network to consider local edge information, where a spatially variant RNN was used to help preserve rich details in the enhanced

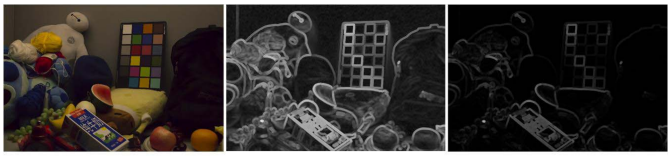


Fig. 2. Comparisons of the contour map obtained by using the EMLV filter with different exponents. The extracted contour map using the EMLV filter with $\gamma = 0.5$ retains more fine-scale details, and the result generated using the EMLV filter with $\gamma = 2$ contains the main shaper edges of the input image.

image. Zhu *et al.* [48] proposed a two-stage network, which first constructs an accurate normal-light image by combining well-exposed areas generated from synthesized multi-exposure images, and then uses edge information to refine the final result. Xu *et al.* [49] introduced a frequency-based method that first recovers image objects in the low-frequency layer and subsequently restores the high-frequency details based on the recovered image objects. In addition to supervised learning methods, several unsupervised methods have been proposed for image enhancement tasks. Jiang *et al.* [51] presented a highly effective unsupervised generative adversarial network trained with unpaired low/normal-light images and proved that it can be generalized well on various real-world test images. In [52], Zhao *et al.* presented a unified deep framework, termed RetinexDIP, which learns the illumination and reflection components without using any external images. Li *et al.* [53] reformulated the enhancement task as a deep curve estimation problem and proposed the first network that can be trained with only input low-light images. Ni *et al.* [54] used a single deep GAN with modulation and attention mechanisms to capture richer global and local features, helping recover normal-light images with unpaired data. To suppress noise, Xiong *et al.* [55] presented a two-stage GAN-based framework to implement illumination enhancement and noise suppression in an unsupervised manner. These unsupervised methods have successfully reduced the effort required to acquire paired low/normal-light images. However, the enhancement performance of these methods cannot compete with that of the supervised learning methods.

The aforementioned CNN-based methods achieved dramatic improvements over traditional methods. However, most CNNs exploit the global structure and local detailed texture in the same way without fully considering the characteristics of these two different components. In this work, to fully consider the global structure and local texture information, we propose a method that utilizes structure-attention and texture-attention subnetworks to separately process structure and texture maps. Moreover, to further explore the internal correlations among global and local features, a fusion subnetwork with channel and spatial attention mechanisms [56] is adopted to integrate the structure and texture information.

III. Methodology

In this section, we present the details of the proposed structure-texture aware network. The architecture of the proposed network is shown in Fig. 1. Our network consists of two parts: a structure-texture decomposition module and a structure-texture aware (STA) module. The structure-texture decomposition module utilizes a contour map-guided filter to decompose the image into structure and texture components. The formation of the decomposition module can be modeled as:

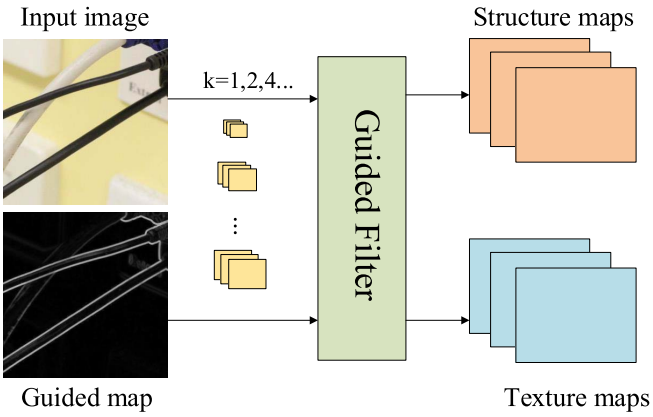


Fig. 3. The structure of the guided decomposition module.

$$S = G(R, C), \tag{1}$$

$$T = R - S, \tag{2}$$

where G is the proposed contour map-guided filter, R is the low-light input image, C is the extracted contour map, S denotes the generated structure map, and T denotes the texture map.

Subsequently, STA module is utilized to generate an enhanced image by integrating the separately extracted global structure features and local texture maps. The STA module can be modeled as:

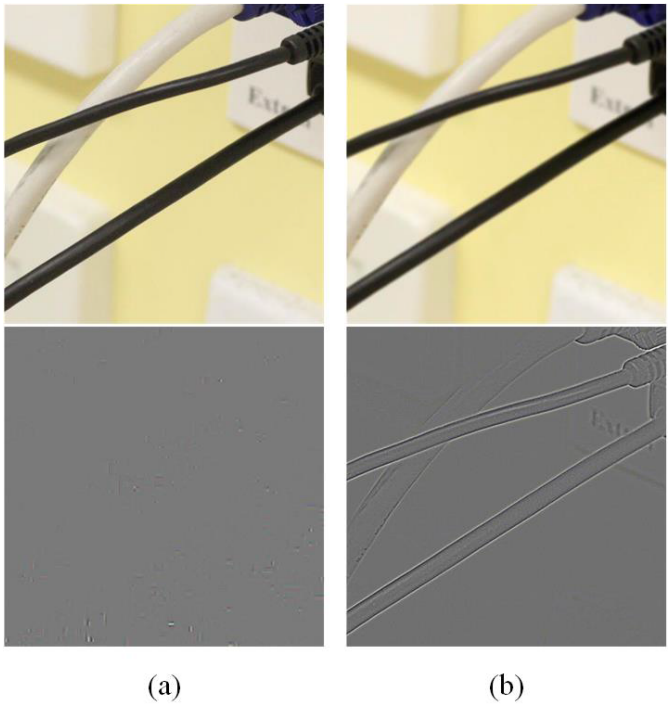
$$O = F(E_N(S), X_N(T)), \tag{3}$$

where F is the fusion subnetwork, E_N is the structure-attention subnetwork, and X_N is the texture-attention subnetwork.

Finally, a hybrid loss function that includes the mean squared error (MSE) loss, total variation (TV) loss, structural dissimilarity (SSIM) loss and color loss is proposed to optimize STANet.

A. Structure-Texture Decomposition Module

The structure-texture decomposition module is designed to decompose the input low-light image into structure maps and texture maps. Inspired by [57], we present a guided decomposition method that uses a contour map for image guidance to generate reasonable structure and texture components. The guidance map plays an important role in the decomposition module. In this work, we adopt a typical mean local variance (MLV) filter [58] to preliminarily estimate the contour map of the scene. Moreover, we utilize exponentiated local derivatives [59], which are generated by introducing an



(a) (b)

Fig. 4. Comparisons of the generated structure map and texture map pairs by using different guided maps. (a) The guided map is the original image. (b) The guided map is our generated contour map ($\gamma = 2$ and $k = 4$).

exponent γ to the local derivatives, to extract the contour map in a global manner.

The typical filter that uses the MLV can be described as:

$$f_{MLV}(R) = \left| \frac{1}{|\Omega|} \sum_{\Omega} \nabla R \right|, \tag{4}$$

where Ω is the local patch around each pixel of R and has a size of 3×3, ∇R denotes the gradient of R. To increase the flexibility of the filter, we add an exponent γ to the MLV filter (EMLV) [59], which can be formally written as:

$$f_{EMLV}(R) = \left| \frac{1}{|\Omega|} \sum_{\Omega} \nabla R \right|^{\gamma}. \tag{5}$$

As illustrated in Fig. 2, γ can help the MLV filter flexibly extract the contours of the input image. When given a large exponent (e.g., $\gamma = 2$), the contour map retains the shaper global edges of the whole image. In contrast, when given a relatively small exponent (e.g., $\gamma = 0.5$), the contour map includes more fine-scale details. Thus, we can extract the contour map flexibly by considering different exponents according to the illumination intensity of the input image.

After obtaining the contour map, we use it to guide the filter and decompose the input image into global structure maps and local texture maps. As shown in Fig. 3, we use a differentiable decomposition layer [60] to design the guided filter, where k represents the smoothing kernel size of the guided filter.

The guided filter allows us to keep the global contour in the structure maps, which contributes to preventing the over-smoothing of the enhanced image. By using the guided filter, we generate structure maps that retain the global structure. Then, texture maps are generated by subtracting the global structure maps from the original image. Fig. 4 illustrates a comparison of the results when different guide maps are used to generate the structure map and texture map. As shown in Fig. 4 (a), the texture map generated by using the input image as a guidance map leads to the loss of basic texture. Compared with the input image-guided filter, the proposed contour-guided filter generates a structure map that retains global structures with shaper edges and a texture map that contains fine details.

B. Structure-Texture Aware Module

The STA module consists of a structure-attention subnetwork, a texture-attention subnetwork, and a fusion subnetwork. The details of the subnetworks are presented as follows.

1) Structure-Attention Subnetwork: To obtain the structure information from an input low-light image, the network requires a large receptive field. Therefore, in this work, we adopt an encoder-decoder architecture that is modified from U-Net [61] to construct the structure-attention subnetwork. Specifically, as shown in Fig. 1, we first feed the decomposed structure maps into a 1×1 convolutional layer, which behaves as a channel-wise feature selector [57]. To fully exploit the extracted features, cascaded residual blocks are used as the feature extractor after each downsampling or upsampling operation. In this way, the proposed structure-attention subnetwork extracts rich global features that contain abundant structure information. Specifically, the structure-attention subnetwork has 21 layers, and the channel number for each convolution layer is set to 64.

2) Texture-Attention Subnetwork: To restore the local details of the low-light image, the network should avoid downsampling operations, which tend to destroy spatial details. Therefore, in this work, we adopt an end-to-end network architecture without any downsampling operations to construct the texture-attention subnetwork. Moreover, to retain the detailed information of the input texture maps, multiple skip connections with addition operations are used. As shown in Fig. 1, the texture-attention subnetwork has 7 layers, and the channel number for each convolutional layer is set to 64. The feature extractor is formed by cascaded convolutional blocks. The leaky ReLU (LReLU) and the instance normalization (IN) are used in each block. Finally, a convolution layer is used to generate the final local features.

3) Fusion Subnetwork: After we extract the global and local features from the structure and texture maps, an integrated fusion network is introduced to integrate these features and reconstruct the final enhanced image. Specifically, channel and spatial attention mechanisms [56] are utilized to discover the internal correlations among the global and local features. The channel attention mechanism adopts a squeeze operation and an excitation operation [62] to capture the relationships among the feature channels, and the spatial attention mechanism applies global average pooling and max pooling along channel dimensions to extract two spatial representation maps; these two maps are then fed into a convolution layer with

TABLE I
AVERAGE PSNR, SSIM, AND FSIM OF ENHANCED RESULTS OBTAINED FROM THE METHODS TESTED ON THE LOL DATASET. THE BEST RESULTS ARE SHOWN IN BOLD FONT

Metrics	BIMEF [63]	CRM [31]	PLM [34]	Dong [64]	LIME [7]	MF [65]	RRM [36]	SRIE [8]
PSNR	13.88	17.20	16.26	16.72	16.76	18.79	13.88	11.86
SSIM	0.5771	0.6442	0.5670	0.5824	0.5644	0.6422	0.6577	0.4979
FSIM	0.9263	0.9442	0.8171	0.8886	0.7698	0.9236	0.8820	0.8390
Metrics	RetinexNet [66]	MSR [33]	NPE [67]	TLLIO [68]	KinD [69]	RetinexDIP [52]	ZeroDCE++ [53]	Ours
PSNR	16.77	13.17	16.97	19.50	20.87	12.24	14.86	**21.35**
SSIM	0.5594	0.4787	0.5894	0.7120	0.8022	0.5872	0.7345	**0.8157**
FSIM	0.7823	0.7060	0.8960	0.8973	0.9397	0.7771	0.9183	**0.9453**

a sigmoid activation function to generate a spatial attention map. Subsequently, the cascaded convolutional layers are used to integrate the output features of the channel attention and spatial attention layers. Finally, a residual learning strategy is adopted to generate the final enhanced image.

C. The Definition of the Hybrid Loss Function

The loss function has a considerable impact on the quality of an enhanced image. In this work, to improve perception quality, a hybrid loss function that includes MSE loss, TV loss, SSIM loss, and color loss is used to train STANet. Formally, the hybrid loss function can be described as:

$$L_{Total} = \lambda_1 L_{MSE} + \lambda_2 L_{TV} + \lambda_3 L_{SSIM} + \lambda_4 L_{Color}, \quad (6)$$

where L_{MSE} indicates the MSE loss, L_{TV} indicates the TV loss, L_{SSIM} denotes the SSIM loss, L_{Color} denotes color loss, and the weights, including λ_1, λ_2, λ_3, and λ_4, are used to balance these losses.

MSE loss [62] is used to ensure the overall content similarity between the enhanced and ground truth images, which is defined as:

$$L_{MSE} = \frac{1}{N}\sum_{i=1}^{N} ||H(R) - I||^2, \quad (7)$$

where N denotes the number of images in each process, $H(R)$ denotes the enhanced image, and I denotes the ground truth image.

TV loss [58] is used to prevent the oversmoothing of the obtained results and is defined as:

$$L_{TV} = \frac{1}{N}\sum_{i=1}^{N} ||\nabla H(R) - \nabla I||^2, \quad (8)$$

where ∇ denotes the gradient operator.

SSIM loss [8] is used to improve the visual effects of the final results and is defined as:

$$L_{SSIM} = \frac{1}{N}\sum_{i=1}^{N} ||SSIM(H(R), I)||^2. \quad (9)$$

Color loss is used to overcome the color distortion problem [70]. Inspired by the definition of color saturation, we define color loss as follows:

$$L_{Color} = \frac{1}{N}\sum_{i=1}^{N} \left\| \frac{max(H(R)) - min(H(R))}{min(H(R))} - \frac{max(I) - min(I)}{min(I)} \right\|^2, \quad (10)$$

where $max(.)$ returns the maximum value of the RGB channels and $min(.)$ returns the minimum value of the RGB channels.

Color saturation affects the visual quality of an enhanced image. However, the MSE loss only measures the overall color difference, and color consistency in each channel is not considered. With the proposed color loss function, the enhanced image achieves improved color consistency in RGB channels and, consequently, enhanced visual quality.

IV. EXPERIMENTAL RESULTS

A. Experimental Settings

1) Implementation Details: The proposed network is implemented using the TensorFlow framework on a single GPU (Tesla V100). The weights are initialized by adopting the scheme proposed in [37]. In addition, the initial learning rate is set to 0.0005, and the decay factor of the learning rate is set to 10 for every 4,000 epochs. During training, the patch size is set to 100 × 100, and the Adam optimizer [69] is adopted to optimize the model with a batch size of 32. For the hybrid loss function, the hyperparameters λ_1, λ_2, λ_3, and λ_4 are empirically set to 1, 0.01, 1, and 0.05, respectively.

2) Dataset Description: We adopt seven popular public datasets, including LOL [66], DPED [9], LIME [7], NPE [67], MEF [71], VV,[1] and DICM [72], to evaluate the performance of the proposed method. The LOL and DPED datasets are widely used for low-light image enhancement tasks, which have corresponding ground truth images. Specifically, the LOL dataset includes 500 low/normal-light image pairs. Following the strategy proposed in [69], 485 image pairs are randomly selected as the training dataset, and the remaining 15 image pairs are used as the test datasets. The DPED dataset contains three sub-datasets captured by different smartphones, including a BlackBerry dataset, an iPhone dataset, and a

[1] https://sites.google.com/site/vonikakis/datasets

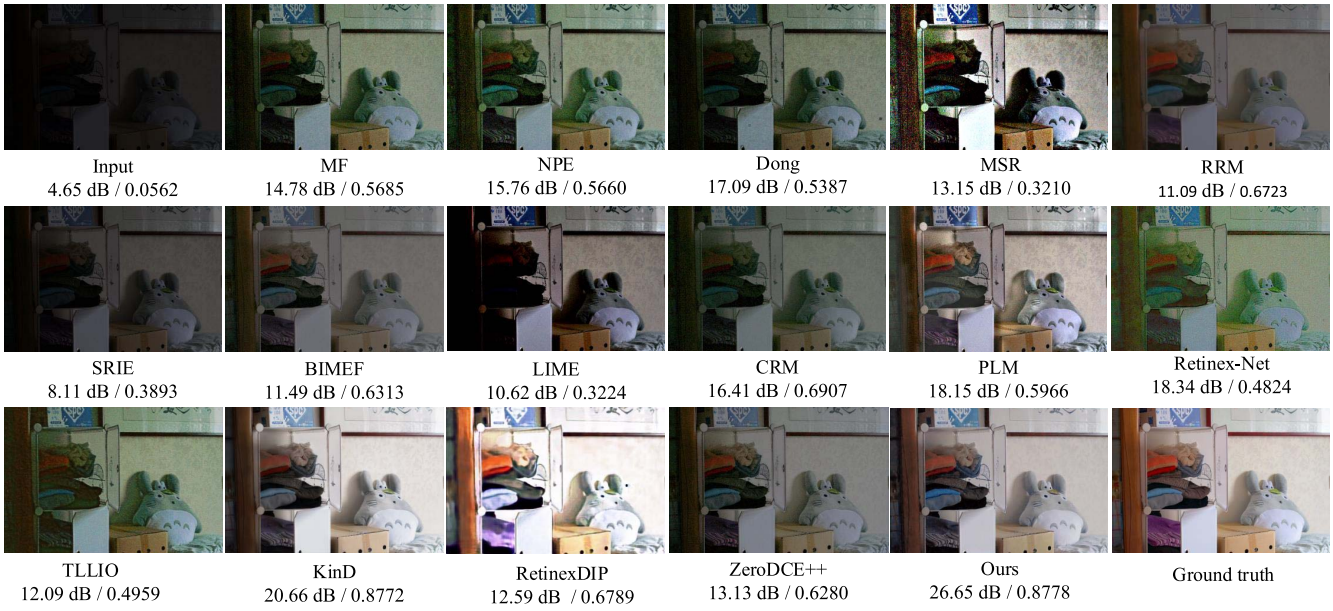


Fig. 5. Qualitative comparison of the results of different methods tested on the LOL dataset. Note the color of the clothes. The proposed method recovers colorful clothes and generates natural illumination in the background.

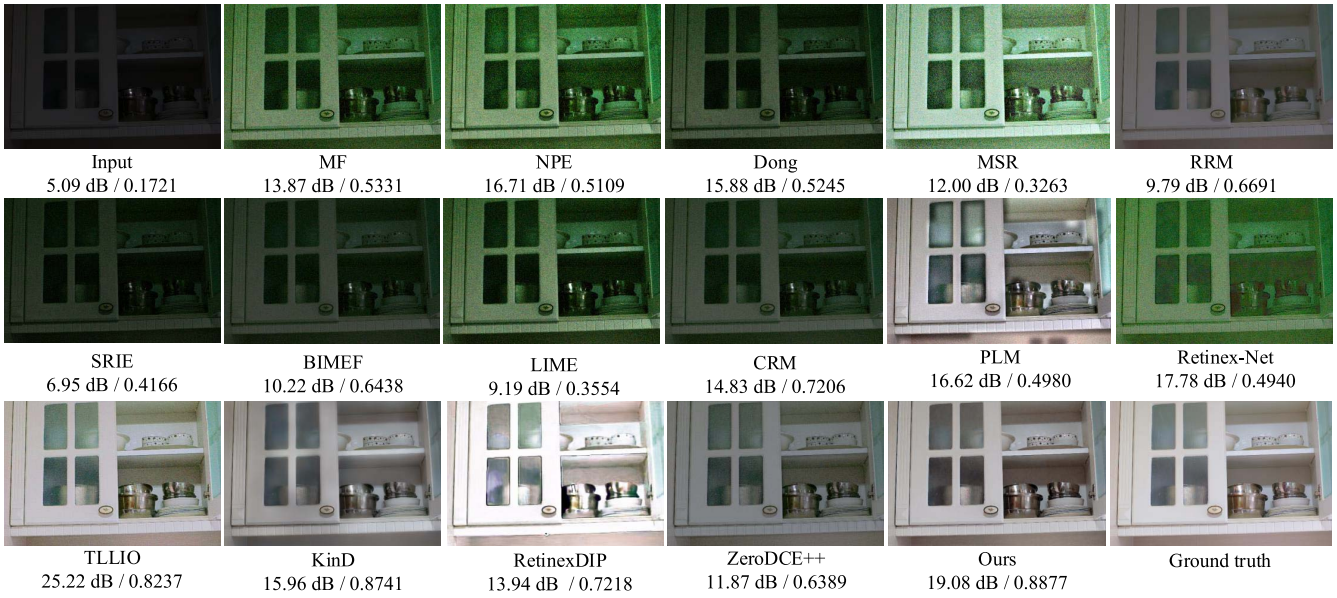


Fig. 6. Qualitative comparison of the results of different methods tested on the LOL dataset. Note the texture of the cupboard. The proposed method recovers more details.

Sony dataset. The corresponding labels are captured by using a Canon DSLR. Images captured by smartphones are often inhibited by low-light conditions, and the images captured by the Canon DSLR have normal brightness levels. We also test the proposed STANet on five non-reference datasets including the LIME (10 images), NPE (84 images), MEF (17 images), VV (24 images), and DICM (44 images) datasets.

3) Evaluation Metrics: In this work, we adopt the peak signal-to-noise ratio (PSNR), the structural similarity index measure (SSIM), and the feature similarity index measure (FSIM) [73] to evaluate the performance of the proposed method on the datasets that include ground truth images. In addition, for the non-reference datasets, the naturalness image quality evaluator (NIQE) [74] index is used to measure the image quality of the enhanced results.

B. Quantitative and Qualitative Evaluation

We compare the proposed STANet method with other state-of-the-art methods, including BIMEF [63], SRIE [8], CRM [31], PLM [34], Dong [64], LIME [7], MF [65], RRM [36], Retinex-Net [66], TLLIO [68], MSR [33], NPE [67], KinD [69], RetinexDIP [52], and ZeroDCE++ [53]. Note that the CNN-based comparison

Fig. 7. Qualitative comparison of the results of different methods on the LOL dataset. Note the color of the wall, the proposed method recovers a clean surface of the wall.

Fig. 8. Qualitative comparison of the results of different methods on the LOL dataset. Note the color of the wall, the proposed method recovers pleasing results.

methods, including RetinexNet, TLLIO, and KinD, are trained on the same dataset as our method, and thus, we use the pretrained model to test the LOL test dataset. The methods RetinexDIP and ZeroDCE++ are trained on different datasets from our method, thus we use the retrained model to test the LOL test dataset. The higher the values of PSNR, SSIM, and FSIM are, the better the quality of the enhanced images. In contrast, for the NIQE, a lower value reflects a higher quality for an enhanced image.

1) For the LOL Dataset: A comparison of the results is presented in Table I. From Table I, we find that the proposed STANet achieves the best PSNR, SSIM, and FSIM results compared to those of all other state-of-the-art methods. Specifically, compared with these methods, the PSNR of the proposed method is improved by more than 0.4 dB, the SSIM is improved by at least 0.013, and the FSIM value is improved by at least 0.0011.

Visual comparisons between the proposed method and other state-of-the-art methods in different scenarios and light conditions are illustrated in Figs. 5-8, the quantitative results in terms of PSNR and SSIM are also listed in the figures. The input images are captured under low-light condition, and the content of the input images cannot be clearly observed. As shown in Figs. 5-6, the enhanced results obtained using NPE [67], MSR [33], and Retinex-Net [66] are biased towards green, which is inconsistent with the real images.

TABLE II

AVERAGE PSNR AND SSIM OF THE ENHANCED RESULTS OBTAINED FROM THE METHODS TESTED ON THE DPED DATASET. THE BEST RESULTS ARE SHOWN IN BOLD FONT

	PSNR/SSIM/FSIM					
	SRIE [8]	HDRNet [75]	PLSR [76]	DSLR-Q [9]	LLDHN [47]	Ours
iphone	**22.22**/0.7320/0.8177	18.31/0.6432/0.7230	20.32/0.9161/0.9241	20.08/0.9201/0.9339	20.90/0.9129/**-**	20.90/**0.9251/0.9351**
blackberry	**20.87**/0.7311/0.7575	18.08/0.6366/0.6745	20.11/0.9298/0.9357	20.07/0.9328/0.9233	20.62/0.9331/**-**	20.51/**0.9359/0.9383**
sony	19.74/0.7250 /0.7402	18.06/0.7895/0.7915	21.33/0.9434/0.9361	21.81/0.9437/0.9204	**22.32**/0.9357/**-**	22.11/**0.9473/0.9363**

Fig. 9. Qualitative comparison of the results of different methods on the DPED dataset.

The enhanced results recovered by Dong [64], SRIE [8], BIMEF [63], and CRM [31] are still dark. PLM [34] generates black artifacts and amplified noise. RetinexDIP [52] produces overexposed local regions and oversmoothed textures. In contrast, the proposed STANet model works well in these cases and can generate an enhanced image that is close to the normal-light scene and contains abundant details. As shown in Figs. 7-8, SRIE [8], BIMEF [63], LIME [7], and ZeroDCE++ [53] generate dim illumination. MSR [33] and Retinex-Net [66] generate amplified noise. RetinexDIP [52] produces artifacts. KinD [69] produces oversmoothed edges and textures. In contrast, the proposed model recovers fine details and achieves improved visual perception quality compared with other competitors.

TABLE III

QUANTITATIVE COMPARISONS OF THE NIQE BASED ON THE LIME, DICM, VV, NPE, AND MEF DATASETS. THE BEST RESULTS ARE BOLDFACED. NOTE THAT A LOW NIQE VALUE INDICATES GOOD IMAGE QUALITY

Metrics	NIQE [74]					
Datasets	LIME	DICM	VV	NPE	MEF	Average
BIMEF [63]	3.8169	3.3403	2.8076	4.1963	3.3054	3.6939
CRM [31]	3.8546	3.3299	2.6175	3.9220	3.2391	3.5329
PLM [34]	3.7868	3.2218	**2.4175**	3.6453	**3.1434**	3.3368
Dong [64]	4.1549	4.0314	2.7671	4.2629	4.0763	3.9817
LIME [7]	4.0516	4.3180	2.8713	4.1263	4.8437	4.0691
MF [65]	4.0689	3.4143	2.5794	4.1096	3.4522	3.6688
RRM [36]	4.6092	3.9560	3.5028	4.0740	5.0269	4.0888
SRIE [8]	3.7863	4.1854	3.4724	3.9795	3.7629	3.9307
Retinex-Net [66]	4.5977	4.5000	2.6952	4.5674	4.3893	4.2845
MSR [33]	5.7425	5.8060	2.7661	4.8904	6.8233	5.0618
NPE [67]	3.7945	3.3769	2.5245	3.4701	3.5060	3.3419
TLLIO [68]	4.2152	3.2738	2.5128	4.0493	3.3005	3.5908
KinD [69]	4.7632	3.5651	3.0267	3.5296	3.8469	3.5699
RetinexDIP [52]	3.8151	3.3726	2.4758	3.5822	3.6577	3.4025
ZeroDCE++ [53]	3.7690	3.5602	3.2169	3.5908	3.2832	3.5138
Ours	3.9901	**3.1227**	2.8105	**3.4543**	3.6463	**3.3346**

2) For the DPED Dataset: We compare the results of our method with those of several state-of-the-art methods in terms of the PSNR, SSIM, and FSIM. Note that the comparison methods use the same training dataset and test dataset as our method. As shown in Table II, the proposed method yields a dramatic improvement in terms of SSIM, which means that the proposed method achieves improved visual perception quality. Moreover, the PSNR achieved by the proposed method is very competitive with those of the other methods, and STANet outperforms most of the other state-of-the-art methods. PSNR represents the average intensity differences between the enhanced image and the ground truth image; thus, an oversmoothed enhanced image may achieve a high PSNR. However, human visual perception is highly adapted for extracting structural and textural information from a scene [77]. The proposed method separately considers the global structure features and local texture features, which helps retain abundant texture details when extracting structure information. Therefore, the proposed STANet model performs better in terms of SSIM, which suggests that the proposed approach is most suitable for the human visual system. Moreover, because DPED is a

TABLE IV
AVERAGE PSNR AND SSIM OF THE ENHANCED RESULTS ON THE LOL DATASET. THE BEST RESULTS ARE HIGHLIGHTED IN BOLD

Metrics	Ours(W L_{MSE})	Ours(W/o L_{SSIM})	Ours(W/o L_{TV})	Ours(W/o L_{Color})	Ours L_{Total}
PSNR	**22.77**	20.24	20.79	20.51	21.35
SSIM	0.7892	0.7486	0.8144	0.7770	**0.8157**

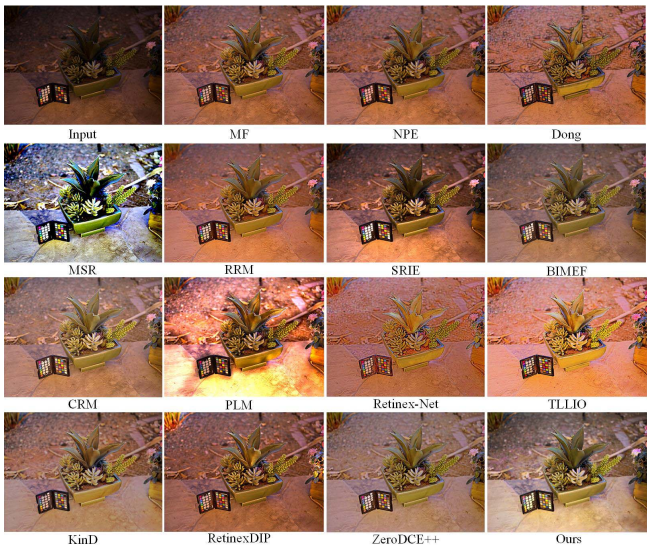


Fig. 10. Qualitative comparison of the results of different methods for a low-light image. Note the global illumination of the image. The proposed method recovers more natural illumination.

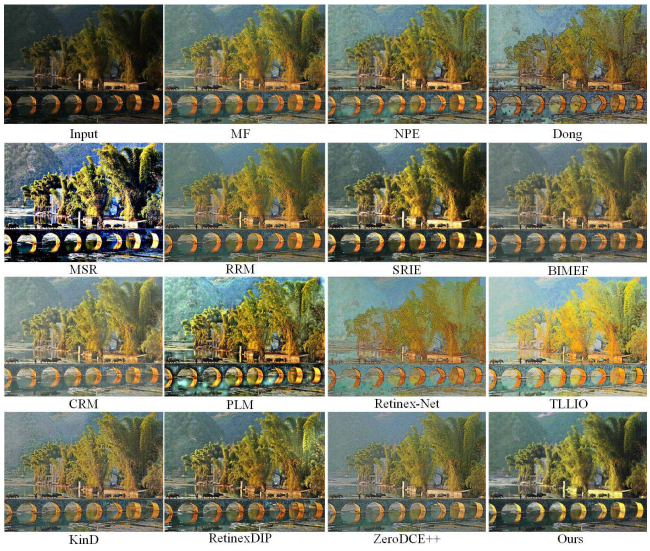


Fig. 11. Qualitative comparison of the results of different methods for a low-light image. Note the color of the beach. The proposed method recovers more natural color.

mobile phone image dataset, the enhanced result requires high visual perception quality. Therefore, although the PSNR result of the proposed method ranks second for some subdatasets, STANet is still very competitive because it achieves better overall visual quality.

Some qualitative comparisons of results are illustrated in Fig. 9. As shown in Fig. 9 (b), SRIE [8] generates an overexposed background. For example, the color of the wall has been changed to white in the first row and the third row. DSLR-Q [9] tends to generate dim regions and artifacts, as shown in Fig. 9 (c). The lightness and color of the images enhanced using the proposed method are more natural than those obtained with other methods. Moreover, as shown in Fig. 9 (d), the proposed method generates more pleasing results in dark regions.

3) For the Non-Reference Datasets: As shown in Table III, the proposed STANet achieves the best results on the NPE and DICM datasets, which decrease the NIQE value by 0.015 and 0.09, respectively. Although the proposed STANet does not obtain the best results on the VV, LIME, and MEF datasets, the average NIQE value of all these datasets is the lowest. These improvements can be attributed to the good generalization ability of the proposed method.

We further present the visual results of the proposed method and other competitors based on these real low-light datasets. As shown in Figs. 10-11, MSR and PLM tend to generate artifacts. Dong, Retinex-Net, and TLLIO produce unnatural illumination. The enhanced results of MF, NPE, RRM, and RetinexDIP have color distortion. In contrast, the proposed model recovers more pleasing results with natural illumination and fine details compared with other competitors.

C. Comparison of Loss Functions

In this section, we discuss the effectiveness of the proposed loss function based on the LOL dataset. The PSNR and SSIM results using L_{MSE}, $L_{MSE} + L_{TV} + L_{Color}$, $L_{MSE} + L_{SSIM} + L_{Color}$, $L_{MSE} + L_{TV} + L_{SSIM}$, and L_{Total} are shown in Table IV. We observe that model using only MSE loss achieves the best performance in terms of the PSNR, which is improved by at least 1.4 dB, but model trained with other losses, including TV loss, SSIM loss, and color loss, can further improve the enhanced visual quality. Especially, the SSIM loss aims to improve the visual quality of the enhanced image. The color loss contributes to improving the color saturation, which makes the enhanced image more colorful. The TV loss tends to alleviate the over smoothing problem so that the results retain abundant image details.

Fig. 12 shows a qualitative comparison of the results of the models trained with L_{MSE}, $L_{MSE} + L_{TV} + L_{Color}$, $L_{MSE} + L_{SSIM} + L_{Color}$, $L_{MSE} + L_{TV} + L_{SSIM}$, and L_{Total}. We can see that all the results are better than the input in Fig. 12 (a). Specifically, the result without SSIM loss looks unnatural as shown in Fig. 12 (c). Removing the TV loss lead to the failed recovery of edge details, as shown in Fig. 12 (d). Severe color

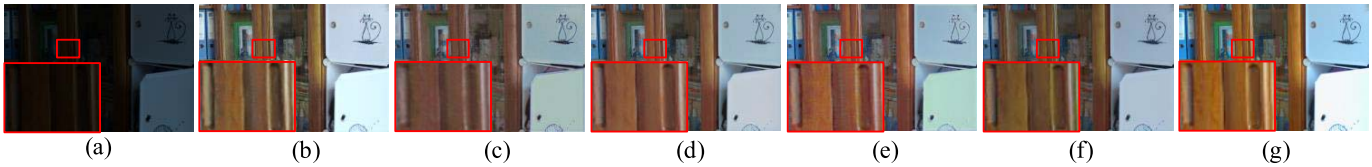


Fig. 12. Visual comparisons of the results of the proposed STANet model trained with different loss functions. (a) Input, (b) L_{MSE}, (c) $L_{MSE} + L_{TV} + L_{Color}$, (d) $L_{MSE} + L_{SSIM} + L_{Color}$, (e) $L_{MSE} + L_{TV} + L_{SSIM}$, (f) L_{Total}, (g) Ground truth.

TABLE V
AVERAGE PSNR AND SSIM OF THE ENHANCED RESULTS ON THE LOL DATASET. THE BEST RESULTS ARE HIGHLIGHTED IN BOLD

Metrics	W/o decomposition	With TV-L1 [16]	With RTV [18]	With SGTD [78]	With RGF [79]	With STDN [28]	With USID [80]	With an input image-guided filter [57]	Ours
PSNR	20.23	17.04	16.94	17.67	21.03	19.69	18.40	20.69	**21.35**
SSIM	0.8036	0.7011	0.7011	0.7226	0.7661	0.7354	0.7219	0.8086	**0.8157**

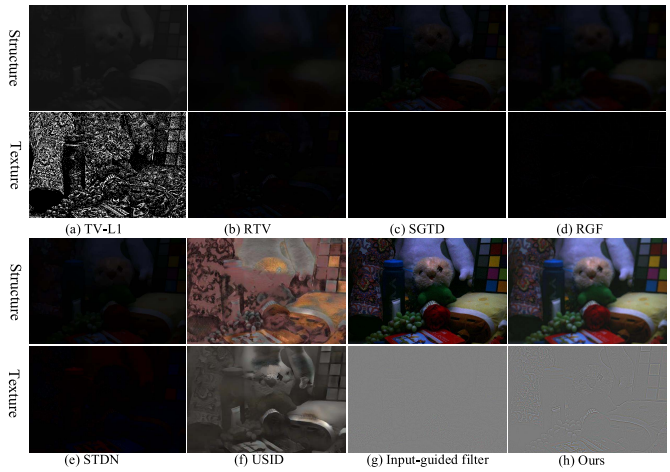


Fig. 13. Visual comparisons of the results of different decomposition methods. (a) TV-L1 [16], (b) RTV [18], (c) SGTD [78], (d) RGF [79], (e) STDN [28], (f) USID [80], (g) Input-guided filter [57], and (h) Ours.

distortion occurs when color loss is discarded, as shown in Fig. 12 (e). The final result using the hybrid loss function is closest to the ground truth image, as shown in Fig. 12 (f).

D. Discussion of the Decomposition Module

To demonstrate the superiority of the adopted decomposition method, we first compare the proposed decomposition module with different decomposition methods. As shown in Fig. 13 (b), (c), (d), and (e), the decomposition methods including RTV [18], SGTD [78], RGF [79], and STDN [28] generate ineffective texture maps that contain amounts of pixels with zero values. As shown in Fig. 13 (f), the method USID [80] generates artifacts. These results can be attributed to the fact that these methods use image priors, such as signed gradients, structure gradients, and repeatability degrees of signal patterns, which may be inapplicable in low-light images, thereby resulting in failed decomposition of structures and textures. In contrast, the proposed decomposition module focuses on low-light images and uses a finer contour map as guided map, which helps generate structure maps that retain global structures with shaper edges and texture maps that contain fine details, as shown in Fig. 13 (h).

Furthermore, we provide the comparison results that apply the decomposition methods to STANet. As shown in Table V, the results show that the decomposition module has a great impact on the final enhancement performance. Compared with our network that does not use the decomposition module, the models using TV-L1 [16], RTV [18], SGTD [78], USID [80], and STDN [28] significantly decrease the enhancement performance, which reduces the PSNR by at least 0.54 dB and the SSIM by at least 0.068. These results occurred because these decomposition methods cannot accurately decompose the low-light images; the decomposed texture maps contain amounts of pixels with zero values. In contrast, the proposed decomposition module helps STANet achieve the best performance, which increases the PSNR and SSIM by at least 0.3 dB and 0.0071, respectively, compared with other competitors. These improvements may have occurred because the proposed decomposition module can effectively extract structure and texture information from the low-light images and thus helps STANet recover more accurate shaper edges and detailed textures in normal-light images. Fig. 14 presents the visual comparison results enhanced by applying different decomposition methods to STANet. As shown in Fig. 14 (b), the recovered result has dim colors when the decomposition module is not used. As shown in Fig. 14 (c), (d), (f), and (g), the results are worse than those in (b). These results occur because the RTV [18], SGTD [78], and USID [80] methods cannot accurately decompose low-light images. The result in Fig. 14 (e) looks more natural and colorful, but the noise is amplified when the RGF method is adopted. The result in Fig. 14 (h) is better than that in (b), but the color still looks unnatural when a decomposition module with an input-guided filter is adopted. In contrast, as shown in Fig. 14 (i), the decomposition module with a contour map-guided filter can be applied to effectively recover edge details and colors. These results indicate that the proposed decomposition module plays a critical role in the proposed method.

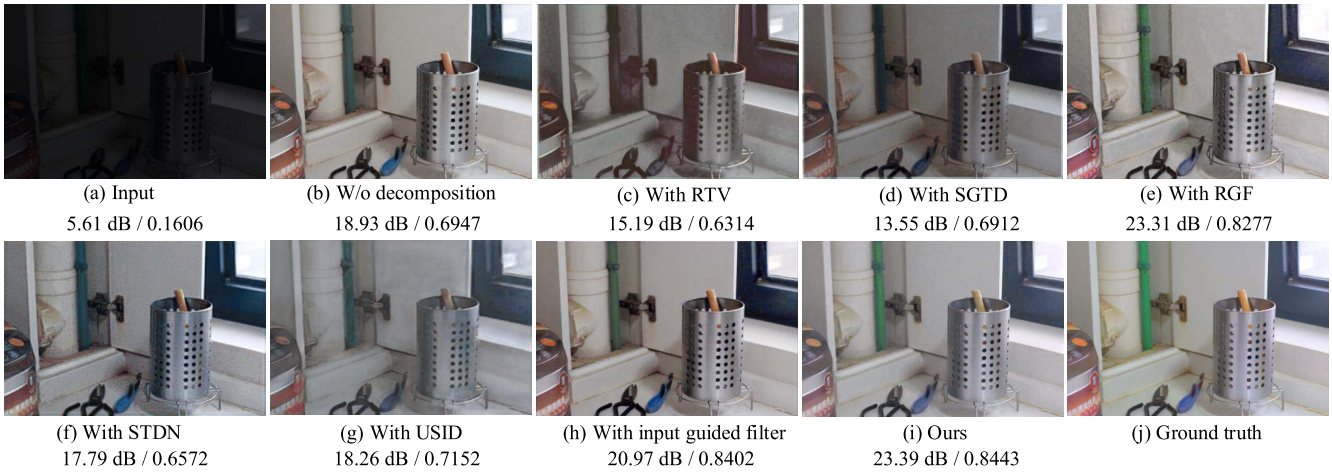


Fig. 14. Visual comparisons of the results of the proposed STANet model trained with different decomposition methods. (a) Input, (b) with TV-L1 [16], (c) with RTV [18], (d) with SGTD [78], (e) with RGF [79], (f) with STDN [28], (g) with USID [80], (h) with input-guided filter [57], (i) Ours, and (j) Ground truth.

TABLE VI
COMPARISON RESULTS OF RUNTIME USING DIFFERENT DECOMPOSITION METHODS

Metrics	TV-L1 [16] (CPU)	RGF [79] (CPU)	RTV [18] (CPU)	SGTD [78] (CPU)	STDN [28] (CPU/GPU)	USID [80] (CPU/GPU)	input-guided filter [57] (CPU/GPU)	Ours (CPU/GPU)
$600 \times 400 \times 3$	0.251 s	24.353 s	2.200 s	9.130 s	1.891 s / 0.154 s	1.565 s / 0.122 s	0.079 s / 0.006 s	0.091 s / 0.007 s
$2080 \times 1560 \times 3$	3.243 s	340.254 s	28.724 s	130.052 s	17.315 s / 1.352 s	15.147 s / 0.941 s	1.212 s / 0.069 s	1.420 s / 0.078 s

To further demonstrate the efficiency of our decomposition module, we compare the running time of our method with other decomposition methods. To ensure fair comparisons, all the decomposition methods are tested on a CPU (Intel Xeon Platinum 8176) or a GPU (Tesla V100). As shown in Table VI, the runtime of the proposed decomposition method is comparable to that of input-guided filter [57], which is more than 10 times faster than that of USID [80] and STDN [28], 20 times faster than that of RTV, 90 times faster than that of SGTD [78], and 230 times faster than that of RGF [79]. This improvement can be attributed to the reason that our method does not require an iterative optimization process.

V. CONCLUSION

The proposed structure-texture decomposition module utilizes a map-guided filter with a fine-scale contour map to decompose a degraded image into structure and texture maps. Global structure maps preserve the overall structure with shaper edges, whereas local texture maps contain fine details. The STA module includes a structure-attention subnetwork, a texture-attention subnetwork, and a fusion subnetwork. The structure-attention subnetwork uses the structure maps as input and adopts an encoder-decoder architecture to extract the global features, which contributes to alleviating artifacts in the enhanced image. In contrast, the texture-attention subnetwork is fed the texture maps and utilizes cascaded residual blocks to process the local features, which helps recover details in the enhanced image. The fusion subnetwork utilizes spatial and channel attention mechanisms to integrate the global and local features and consider the internal correlations among these features. Thus, the proposed method successfully considers structure and texture information, which makes the STANet retain abundant detailed information when extracting global structure information. Moreover, we propose a hybrid loss function to optimize STANet, where a color loss function is utilized to alleviate the color distortion in the enhanced image. Through this approach, the proposed STANet model achieves better visual quality than other state-of-the-art methods.

REFERENCES

[1] P. Tang *et al.*, "PCL: Proposal cluster learning for weakly supervised object detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 1, pp. 176–191, Jan. 2020.
[2] M. Abdullah-Al-Wadud, M. H. Kabir, M. A. A. Dewan, and O. Chae, "A dynamic histogram equalization for image contrast enhancement," *IEEE Trans. Consum. Electron.*, vol. 53, no. 2, pp. 593–600, May 2007.
[3] T. Arici, S. Dikbas, and Y. Altunbasak, "A histogram modification framework and its application for image contrast enhancement," *IEEE Trans. Image Process.*, vol. 18, no. 9, pp. 1921–1935, Sep. 2009.
[4] D. Sen and S. K. Pal, "Automatic exact histogram specification for contrast enhancement and visual system based quantitative evaluation," *IEEE Trans. Image Process.*, vol. 20, no. 5, pp. 1211–1220, May 2011.
[5] C. Lee, C. Lee, and C.-S. Kim, "Contrast enhancement based on layered difference representation of 2D histograms," *IEEE Trans. Image Process.*, vol. 22, no. 12, pp. 5372–5384, Dec. 2013.
[6] H. D. Cheng and X. J. Shi, "A simple and effective histogram equalization approach to image enhancement," *Digit. Signal Process.*, vol. 14, no. 2, pp. 158–170, Mar. 2004.
[7] X. Guo, Y. Li, and H. Ling, "LIME: Low-light image enhancement via illumination map estimation," *IEEE Trans. Image Process.*, vol. 26, no. 2, pp. 982–993, Feb. 2017.
[8] X. Fu, D. Zeng, Y. Huang, X.-P. Zhang, and X. Ding, "A weighted variational model for simultaneous reflectance and illumination estimation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 2782–2790.

[9] A. Ignatov, N. Kobyshev, R. Timofte, and K. Vanhoey, "DSLR-quality photos on mobile devices with deep convolutional networks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 3277–3285.

[10] J. Huang *et al.*, "Range scaling global U-Net for perceptual image enhancement on mobile devices," in *Proc. Eur. Conf. Comput. Vis. (ECCV) Workshops*, Sep. 2018, pp. 230–242.

[11] J. Li, X. Feng, and Z. Hua, "Low-light image enhancement via progressive-recursive network," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 31, no. 11, pp. 4227–4240, Nov. 2021.

[12] J. Song, H. Cho, J. Yoon, and S. M. Yoon, "Structure adaptive total variation minimization-based image decomposition," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 28, no. 9, pp. 2164–2176, Sep. 2018.

[13] J.-F. Aujol, G. Gilboa, T. Chan, and S. Osher, "Structure-texture image decomposition—Modeling, algorithms, and parameter selection," *Int. J. Comput. Vis.*, vol. 67, no. 1, pp. 111–136, Apr. 2006.

[14] L. I. Rudin, S. Osher, and E. Fatemi, "Nonlinear total variation based noise removal algorithms," *Phys. D, Nonlinear Phenomena*, vol. 60, nos. 1–4, pp. 259–268, 1992.

[15] Y. Meyer, *Oscillating Patterns in Image Processing and Nonlinear Evolution Equations: The Fifteenth Dean Jacqueline B. Lewis Memorial Lectures*, vol. 22. Providence, RI, USA: American Mathematical Society, 2001.

[16] W. Yin, D. Goldfarb, and S. Osher, "Image cartoon-texture decomposition and feature selection using the total variation regularized L^1 functional," in *Proc. Int. Workshop Variational, Geometric, Level Set Methods Comput. Vis.* Berlin, Germany: Springer, 2005, pp. 73–84.

[17] J.-F. Aujol and S. H. Kang, "Color image decomposition and restoration," *J. Vis. Commun. Image Represent.*, vol. 17, no. 4, pp. 916–928, 2006.

[18] L. Xu, Q. Yan, Y. Xia, and J. Jia, "Structure extraction from texture via relative total variation," *ACM Trans. Graph.*, vol. 31, no. 6, pp. 1–10, 2012.

[19] Y. Kim, B. Ham, M. N. Do, and K. Sohn, "Structure-texture image decomposition using deep variational priors," *IEEE Trans. Image Process.*, vol. 28, no. 6, pp. 2692–2704, Jun. 2019.

[20] H. Liu, R. Xiong, Q. Song, F. Wu, and W. Gao, "Image super-resolution based on adaptive joint distribution modeling," in *Proc. IEEE Vis. Commun. Image Process. (VCIP)*, Dec. 2017, pp. 1–4.

[21] F. Zhu, C. Fang, and K.-K. Ma, "PNEN: Pyramid non-local enhanced networks," *IEEE Trans. Image Process.*, vol. 29, pp. 8831–8841, 2020.

[22] H. Yin, Y. Gong, and G. Qiu, "Side window filtering," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 8758–8766.

[23] P. Xu and W. Wang, "Structure-aware window optimization for texture filtering," *IEEE Trans. Image Process.*, vol. 28, no. 9, pp. 4354–4363, Sep. 2019.

[24] Y. Ma, X. Jiang, Z. Xia, M. Gabbouj, and X. Feng, "CasQNet: Intrinsic image decomposition based on cascaded quotient network," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 31, no. 7, pp. 2661–2674, Jul. 2021.

[25] K. Lu, S. You, and N. Barnes, "Deep texture and structure aware filtering network for image smoothing," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Sep. 2018, pp. 217–233.

[26] Q. Fan, J. Yang, D. Wipf, B. Chen, and X. Tong, "Image smoothing via unsupervised learning," *ACM Trans. Graph.*, vol. 37, no. 6, pp. 1–14, 2018.

[27] X. Gao, X. Wu, P. Xu, S. Guo, M. Liao, and W. Wang, "Semi-supervised texture filtering with shallow to deep understanding," *IEEE Trans. Image Process.*, vol. 29, pp. 7537–7548, 2020.

[28] F. Zhou, Q. Chen, B. Liu, and G. Qiu, "Structure and texture-aware image decomposition via training a neural network," *IEEE Trans. Image Process.*, vol. 29, pp. 3458–3473, 2020.

[29] Z. Ying, G. Li, Y. Ren, R. Wang, and W. Wang, "A new low-light image enhancement algorithm using camera response model," in *Proc. IEEE Int. Conf. Comput. Vis. Workshops (ICCVW)*, Oct. 2017, pp. 3015–3022.

[30] X. Xu *et al.*, "Rendering portraitures from monocular camera and beyond," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Sep. 2018, pp. 36–51.

[31] Y. Ren, Z. Ying, T. H. Li, and G. Li, "LECARM: Low-light image enhancement using the camera response model," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 29, no. 4, pp. 968–981, Apr. 2018.

[32] E. H. Land, "The Retinex theory of color vision," *Sci. Amer.*, vol. 237, no. 6, pp. 108–128, Dec. 1977.

[33] D. J. Jobson, Z.-U. Rahman, and G. A. Woodell, "A multiscale Retinex for bridging the gap between color images and the human observation of scenes," *IEEE Trans. Image Process.*, vol. 6, no. 7, pp. 965–976, Jul. 1997.

[34] S.-Y. Yu and H. Zhu, "Low-illumination image enhancement algorithm based on a physical lighting model," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 29, no. 1, pp. 28–37, Jan. 2019.

[35] B.-H. Chen, Y.-L. Wu, and L.-F. Shi, "A fast image contrast enhancement algorithm using entropy-preserving mapping prior," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 29, no. 1, pp. 38–49, Jan. 2019.

[36] M. Li, J. Liu, W. Yang, X. Sun, and Z. Guo, "Structure-revealing low-light image enhancement via robust Retinex model," *IEEE Trans. Image Process.*, vol. 27, no. 6, pp. 2828–2841, Jun. 2018.

[37] C. Chen, Z. Xiong, X. Tian, and F. Wu, "Deep boosting for image denoising," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Sep. 2018, pp. 3–19.

[38] H. Chen, Y. Jin, M. Duan, C. Zhu, and E. Chen, "DOF: A demand-oriented framework for image denoising," *IEEE Trans. Ind. Informat.*, vol. 17, no. 8, pp. 5369–5379, Aug. 2021.

[39] H. Chen, Y. Jin, K. Xu, Y. Chen, and C. Zhu, "Multiframe-to-multiframe network for video denoising," *IEEE Trans. Multimedia*, early access, May 3, 2021, doi: 10.1109/TMM.2021.3077140.

[40] Y.-T. Peng, Z. Lu, F.-C. Cheng, Y. Zheng, and S.-C. Huang, "Image haze removal using airlight white correction, local light filter, and aerial perspective prior," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 30, no. 5, pp. 1385–1395, May 2020.

[41] S. Y. Kim, J. Oh, and M. Kim, "Deep SR-ITM: Joint learning of super-resolution and inverse tone-mapping for 4K UHD HDR applications," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 3116–3125.

[42] K. G. Lore, A. Akintayo, and S. Sarkar, "LLNet: A deep autoencoder approach to natural low-light image enhancement," *Pattern Recognit.*, vol. 61, pp. 650–662, Jan. 2017.

[43] L. Shen, Z. Yue, F. Feng, Q. Chen, S. Liu, and J. Ma, "MSR-Net: Low-light image enhancement using deep convolutional network," 2017, *arXiv:1711.02488*.

[44] Q. Chen, J. Xu, and V. Koltun, "Fast image processing with fully-convolutional networks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2497–2506.

[45] J. Liang, Y. Xu, Y. Quan, J. Wang, H. Ling, and H. Ji, "Deep bilateral retinex for low-light image enhancement," 2020, *arXiv:2007.02018*.

[46] J. Liu and C. Jung, "Multiple connected residual network for image enhancement on smartphones," in *Proc. Eur. Conf. Comput. Vis. (ECCV) Workshops*, Sep. 2018, pp. 182–196.

[47] W. Ren, S. Liu, L. Ma, Q. Xu, X. Xu, X. Cao, J. Du, and M.-H. Yang, "Low-light image enhancement via a deep hybrid network," *IEEE Trans. Image Process.*, vol. 28, no. 9, pp. 4364–4375, Sep. 2019.

[48] M. Zhu, P. Pan, W. Chen, and Y. Yang, "Eemefn: Low-light image enhancement via edge-enhanced multi-exposure fusion network," in *Proc. AAAI Conf. Artif. Intell.*, vol. 34, no. 7, 2020, pp. 13106–13113.

[49] K. Xu, X. Yang, B. Yin, and R. W. H. Lau, "Learning to restore low-light images via decomposition-and-enhancement," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 2281–2290.

[50] F. Lv, Y. Li, and F. Lu, "Attention guided low-light image enhancement with a large scale low-light simulation dataset," *Int. J. Comput. Vis.*, pp. 1–19, 2021.

[51] Y. Jiang, X. Gong, D. Liu, Y. Cheng, and C. Fang, "EnlightenGAN: Deep light enhancement without paired supervision," *IEEE Trans. Image Process.*, vol. 30, pp. 2340–2349, 2021.

[52] Z. Zhao, B. Xiong, L. Wang, Q. Ou, L. Yu, and F. Kuang, "RetinexDIP: A unified deep framework for low-light image enhancement," *IEEE Trans. Circuits Syst. Video Technol.*, early access, Apr. 15, 2021, doi: 10.1109/TCSVT.2021.3073371.

[53] C. Li, C. Guo, and C. L. Chen, "Learning to enhance low-light image via zero-reference deep curve estimation," *IEEE Trans. Pattern Anal. Mach. Intell.*, early access, Mar. 3, 2021, doi: 10.1109/TPAMI.2021.3063604.

[54] Z. Ni, W. Yang, S. Wang, L. Ma, and S. Kwong, "Towards unsupervised deep image enhancement with generative adversarial network," *IEEE Trans. Image Process.*, vol. 29, pp. 9140–9151, 2020.

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

[56] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "CBAM: Convolutional block attention module," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Sep. 2018, pp. 3–19.

[57] R. Li, L.-F. Cheong, and R. T. Tan, "Heavy rain image restoration: Integrating physics model and conditional adversarial learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 1633–1642.

[58] B. Cai, X. Xu, K. Guo, K. Jia, B. Hu, and D. Tao, "A joint intrinsic-extrinsic prior model for retinex," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 4020–4029.

[59] J. Xu *et al.*, "STAR: A structure and texture aware retinex model," *IEEE Trans. Image Process.*, vol. 29, pp. 5022–5037, 2020.

[60] H. Wu, S. Zheng, J. Zhang, and K. Huang, "Fast end-to-end trainable guided filter," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 1838–1847.

[61] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent*, 2015, pp. 234–241.

[62] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7132–7141.

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

[64] X. Dong, Y. Pang, and J. Wen, "Fast efficient algorithm for enhancement of low lighting video," in *Proc. ACM SIGGRAPH Posters (SIGGRAPH)*, 2010, pp. 1–6.

[65] X. Fu, D. Zeng, Y. Huang, Y. Liao, X. Ding, and J. Paisley, "A fusion-based enhancing method for weakly illuminated images," *Signal Process.*, vol. 129, pp. 82–96, Dec. 2016.

[66] C. Wei, W. Wang, W. Yang, and J. Liu, "Deep retinex decomposition for low-light enhancement," 2018, *arXiv:1808.04560*.

[67] S. Wang, J. Zheng, H.-M. Hu, and B. Li, "Naturalness preserved enhancement algorithm for non-uniform illumination images," *IEEE Trans. Image Process.*, vol. 22, no. 9, pp. 3538–3548, Sep. 2013.

[68] Y. Zhang, X. Di, B. Zhang, and C. Wang, "Self-supervised image enhancement network: Training with low light images only," 2020, *arXiv:2002.11300*.

[69] Y. Zhang, J. Zhang, and X. Guo, "Kindling the darkness: A practical low-light image enhancer," in *Proc. 27th ACM Int. Conf. Multimedia*, Oct. 2019, pp. 1632–1640.

[70] R. Liu, X. Fan, M. Zhu, M. Hou, and Z. Luo, "Real-world underwater enhancement: Challenges, benchmarks, and solutions under natural light," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 30, no. 12, pp. 4861–4875, Dec. 2020.

[71] K. Ma, K. Zeng, and Z. Wang, "Perceptual quality assessment for multi-exposure image fusion," *IEEE Trans. Image Process.*, vol. 24, no. 11, pp. 3345–3356, Nov. 2015.

[72] C. Lee, C. Lee, and C.-S. Kim, "Contrast enhancement based on layered difference representation," in *Proc. 19th IEEE Int. Conf. Image Process.*, Sep. 2012, pp. 965–968.

[73] L. Zhang, L. Zhang, X. Mou, and D. Zhang, "FSIM: A feature similarity index for image quality assessment," *IEEE Trans. Image Process.*, vol. 20, no. 8, pp. 2378–2386, Aug. 2011.

[74] A. Mittal, R. Soundararajan, and A. C. Bovik, "Making a 'completely blind' image quality analyzer," *IEEE Signal Process. Lett.*, vol. 20, no. 3, pp. 209–212, 2013.

[75] M. Gharbi, J. Chen, J. T. Barron, S. W. Hasinoff, and F. Durand, "Deep bilateral learning for real-time image enhancement," *ACM Trans. Graph.*, vol. 36, no. 4, p. 118, Jul. 2017.

[76] J. Johnson, A. Alahi, and L. Fei-Fei, "Perceptual losses for real-time style transfer and super-resolution," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2016, pp. 694–711.

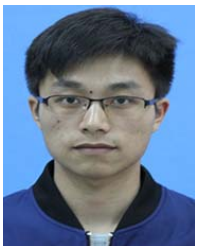
[77] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, Apr. 2004.

[78] Q. Liu, J. Liu, P. Dong, and D. Liang, "SGTD: Structure gradient and texture decorrelating regularization for image decomposition," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2013, pp. 1081–1088.

[79] Q. Zhang, X. Shen, L. Xu, and J. Jia, "Rolling guidance filter," in *Proc. Eur. Conf. Comput. Vis.*, Sep. 2014, pp. 815–830.

[80] Y. Liu, Y. Li, S. You, and F. Lu, "Unsupervised learning for intrinsic image decomposition from a single image," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 3248–3257.

Kai Xu received the B.S. degree in electronic and information engineering from the Anhui University of Science and Technology, Huainan, China, in 2013. She is currently pursuing the Ph.D. degree in precision machinery and precision instruments with the School of Engineering Science, University of Science and Technology of China, Hefei. Her research interests include deep learning and image enhancement.



Huaian Chen received the B.S. degree in mechanical design, manufacturing, and automation from Anhui University, Hefei, China, in 2017. He is currently pursuing the Ph.D. degree in precision machinery and precision instruments with the School of Engineering Science, University of Science and Technology of China, Hefei. His research interests include deep learning, image/video segmentation, and image/video restoration.

Chunmei Xu received the B.S. degree in engineering mechanics from Xinjiang University, Urumqi, China, in 2011. She is currently pursuing the Ph.D. degree in precision machinery and precision instruments with the School of Engineering Science, University of Science and Technology of China, Hefei. She has been working as a Lecturer with the Southwest University of Science and Technology since 2011. Her research interests include image processing and 3D shape measurement.

Yi Jin (Member, IEEE) received the Ph.D. degree from the University of Science and Technology of China, Hefei, China, in 2013.

He is currently an Associate Professor with the School of Engineering Science, University of Science and Technology of China. He has authored or coauthored over 50 refereed articles. His current research interests include intellectual detection, image processing, and artificial intelligence. He was a recipient of the Key Innovations Award from the Chinese Academy of Sciences and the First Class Science and Technology Progress Award in Anhui in 2016 and 2019, respectively.

Changan Zhu received the Ph.D. degree from the National University of Defense Technology, Changsha, China, in 1989.

He is currently a Professor with the School of Engineering Science, University of Science and Technology of China, Hefei, China. He has authored or coauthored over 100 refereed articles. His general areas of research include signal processing, control theory, and intelligent manufacture. He was a recipient of the First Class Award for Scientific and Technological Progress of National Defense and the Second Class Technological Invention Award from the Chinese Institute of Electronics in 2007 and 2019, respectively.