Core-attributes enhanced generative adversarial networks for robust image enhancement

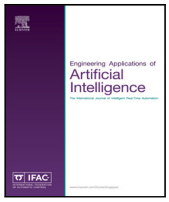
Liu, S.; Xiao, G.; Lew, M.S.K.; Gao, X.; Wu, S.

Engineering Applications of Artificial Intelligence 131 (2024) 107799

Contents lists available at ScienceDirect

Engineering Applications of Artificial Intelligence

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



Research paper

Core-attributes enhanced generative adversarial networks for robust image enhancement

Shan Liu [a], Guoqiang Xiao [a], Michael S. Lew [b], Xinbo Gao [c], Song Wu [a,*]

[a] College of Computer and Information Science, Southwest University, Chongqing, China
[b] Liacs Media Lab, Leiden University, Leiden, Netherlands
[c] College of Computer Science and Technology, Chongqing University of Posts and Telecommunications, Chongqing, China

ARTICLE INFO

Keywords:
Image enhancement
Image retouching
Generative adversarial network
Unsupervised learning

ABSTRACT

Automated image enhancement algorithms have a profound impact on human life today. To solve the problems of luminance, lack of detail information, and overall color tone bias of images taken by mobile devices, a novel framework of core-attributes enhanced generative adversarial network (CAE-GAN) is designed to improve these core attributes of enhanced images. The generator in CAE-GAN mainly consists of a luminance correction encoder (LCE) and a high-frequency supplementary decoder (HFSD). To target the adaptive luminance improvement for each location, the encoder based on LCE is designed by combining the extracted prior knowledge of luminance. Meanwhile, a decoder based on HFSD is proposed to fill in missing edge details during the image reconstruction process. In addition, a multi-scale statistical characteristics distinction branch (MSCDB) is proposed to correct the overall tone. Moreover, an upgrade adversarial loss function is designed to focus on the discrimination of both multi-scale and multi-perspective. The generator and discriminator are iteratively trained under the constraints of the total loss function, resulting in the generator that automatically improves the visualization of the images. Extensive experiments have shown that our CAE-GAN is capable of achieving excellent results in several evaluation metrics and subjective results. The source code of the proposed CAE-GAN is available at https://github.com/SWU-CS-MediaLab/CAE-GAN.

1. Introduction

It is popular to use mobile phones to record our daily life. However, external factors, such as illumination, weather and climate, and even the quality of the device, may affect the quality of the captured pictures. Most people would choose to edit the photos before sharing them. However, this is more laborious for ordinary users due to the need for proficient retouching skills and aesthetics compared to professional photographers. Therefore, automatic image retouching is urgently required to bridge the gap with professional image enhancement and improve public convenience.

Traditional image enhancement algorithms still suffer some limitations. In previous studies, histogram equalization-based (Pizer et al., 1987) and its variants (Lee et al., 2013), and Retinex-based methods (Li et al., 2018) have been able to modify the quality of images to some extent. However, those methods cannot focus on the full range of image properties, which results in specific weaknesses in the final image enhancement quality, such as over-enhancing luminance or contrast, losing local details, and failing to adapt to different scenes, etc.

Deep neural networks (DNNs) (Xu et al., 2022; Luo et al., 2023; Zou et al., 2022) have recently demonstrated their superior capability in various computer vision tasks. Thus, improving the quality of images by optimizing the image enhancement object functions based on deep neural networks is gaining increasing attention. In terms of achieved effects, the image enhancement task can first be divided into low-light image enhancement (Jiang et al., 2021; Guo et al., 2020) and image retouching (Ni et al., 2020; Liu et al., 2022; He et al., 2020). The former was valued earlier and only focused on enhancing the light effects alone, while the latter began to appear gradually only after people had new demands on aesthetics. Each image retouching method has obtained visually varying results due to differences in algorithm design, and they have more in common with an emphasis on saturation and global tone. This further demonstrates the significance of light, details, and color tones for the high quality of image enhancement.

Moreover, most existing image enhancement methods train and optimize based on paired datasets, but it is so demanding (pixel-by-pixel alignment) that it takes time to obtain. The most widely used

* Corresponding author.
E-mail addresses: swu1538@email.swu.edu.cn (S. Liu), gqxiao@swu.edu.cn (G. Xiao), m.s.lew@liacs.leidenuniv.nl (M.S. Lew), gaoxb@cqupt.edu.cn (X. Gao), songwuswu@swu.edu.cn (S. Wu).

https://doi.org/10.1016/j.engappai.2023.107799
Received 22 May 2023; Received in revised form 28 November 2023; Accepted 22 December 2023
Available online 29 December 2023

Fig. 1. The proposed CAE-GAN achieves visually pleasing performance on low-quality images. Low-quality images have problems of darkness, low clarity, and lack of color. Our proposed CAE-GAN can effectively transfer the low-quality domain into the high-quality domain and significantly improve its visual impact.

datasets are manually edited by professional photographers one by one, which is a time-consuming and tedious process (Bychkovsky et al., 2011). Hence, we aspire to explore unsupervised image enhancement methods using unpaired datasets.

According to previous research, it is found that image enhancement can be considered an unpaired image-to-image translation task (Isola et al., 2017), which gradually focuses on and extends to the field of image enhancement. However, traditional models only accomplish coarse translation on contours and cannot satisfy the requirement of brightness, saturation, and full-image tone in high-quality image enhancement. The laplacian pyramid translation network (LPTN) (Liang et al., 2021) achieved fine conversion of high-resolution images. However, since it is not precise enough as a general-purpose framework, the LPTN is not ideal for image enhancement. Since the generative adversarial networks (GANs) (Goodfellow et al., 2014) can learn the mapping relationships between cross-domain, it is believed that the GANs benefit in exploring the cross-domain translation between both low- and high-quality domains. The unsupervised image enhancement GAN (UEGAN) (Ni et al., 2020) used statistical information to assist the GANs in enhancing the image quality, in which the used statistical information was considered to be useful in capturing aesthetic knowledge. However, only global information is utilized in UEGAN, resulting in artifacts and low saturation problems. The bi-directional normalization and color attention-guided GAN (BNCAGAN) (Liu et al., 2022) proposed a bi-directional normalization generator that guarantees details, saturation, and high aesthetic tones, but suffers a bit dark in terms of brightness. Therefore, balancing brightness, saturation, full-image tone, and image details is a challenging and worthwhile exploration of the task of high-quality image enhancement.

Thus, in this paper, a novel framework of core-attributes enhanced generative adversarial network (CAE-GAN) is designed to comprehensively hit the core points overlooked in most existing methods, including illumination, fine details, overall color tone, and domain style. First, the image enhancement process is abstracted as a mapping from low-quality to high-quality domains, and the GAN is adopted as the main framework to focus on learning the precise and robust mapping functions. An effective luminance correction encoder (LCE) is designed using the extracted luminance information of the input images as prior knowledge to enhance illumination effects. Meanwhile, a high-frequency supplementary decoder (HFSD) is designed to complement the high-frequency details lost in the encoding process by extracting high-frequency information from the laplacian pyramid (LP), which is necessary for enhancing the fine local details. Moreover, inspired by the style transfer (Huang and Belongie, 2017), statistical characteristics can represent the style of an image in the generator, and most existing methods align or fusion of target domain style statistical features in the generator.

However, on the one hand, the align or fusion process results in too many constraints and strengthens the complexity and instability of the training process in the generator; on the other hand, low-level statistical characteristics represent global stability, while high-level ones are highly semantic discriminative, and both of them are useful for the discriminator. Thus, designing an adversarial manner in the discriminator based on low-level and high-level statistical characteristics in a multi-scale fusion manner is more reasonable. The multi-scale statistical characteristics distinction branch (MSCDB) is designed to judge the enhancement effect. Furthermore, the MSCDB is combined with the color attention module (CAM) (Liu et al., 2022) to correct the overall tone. The excellent image enhancement results can be finally achieved by CAE-GAN as shown in Fig. 1.

To summarize, the main contributions of our proposed CAE-GAN are summarized as follows:

- A novel and effective framework of core-attributes enhanced generative adversarial network (CAE-GAN) is proposed for high-quality image enhancement. The CAE-GAN consists of the luminance correction encoder (LCE), the high-frequency supplementary decoder (HFSD), and the multi-scale statistical characteristics distinction branch (MSCDB), and mainly targets the core attributes enhancement effects, such as brightness, fine details, overall color tone, and domain style.
- A luminance correction encoder (LCE) is designed to overcome the darkness visibility by extracting the prior luminance knowledge. A high-frequency supplementary decoder (HFSD) is proposed to supervise the fine local details of the reconstructed images by the high-frequency information of the established laplacian pyramid (LP). The generator in CAE-GAN can be effectively optimized by combining both the structure of LCE and HFSD.
- For the discriminator in CAE-GAN, a multi-scale statistical characteristics distinction branch (MSCDB) is designed by considering low-level and high-level statistical characteristics in a multi-scale manner, which can effectively judge the global and local style of images to ensure that the enhanced image quality is consistent with the target domain.
- The experimental results on the MIT-Adobe FiveK dataset and DPED dataset demonstrated the superiority of our proposed CAE-GAN quantitatively and qualitatively, and the CAE-GAN also achieves competitively or even better enhancement performance compared to the other methods.

Section 2 reviewed the generative adversarial networks, image enhancement, and laplacian pyramid. Section 3 presents the details of the proposed CAE-GAN framework and loss function. In Section 4, the experimental setup is introduced. Meanwhile, the results of comparison and ablation experiments are shown in terms of metrics and visual effects. The advantages as well as limitations of the method are explored in Section 5. Finally, we summarize the entire paper in Section 6.

2. Related work

2.1. Generative adversarial network

Generative adversarial networks (GANs) (Goodfellow et al., 2014) have been widely used in various image vision fields since it was proposed and has achieved outstanding performance. Radford et al. in Radford et al. (2015) combined GANs with deep neural networks to make good progress in the image generation task. Ledig et al. in Ledig et al. (2017) applied GANs to the task of image super-resolution, which is a GAN-based network optimized by a novel perceptual loss. Isola et al. in Isola et al. (2017) designed an image-to-image translation model based on U-Net and conditional GAN to handle paired datasets based image-to-image translation task, while Zhu et al. in Zhu et al. (2017) effectively solved the image-to-image translation task by cycle consistency loss and two-way GAN structure based on the unpaired datasets. Zhang et al. in Zhang et al. (2019) introduced a self-attention mechanism to improve focus on key areas. In addition, least squares GAN (LSGAN) (Mao et al., 2017), wasserstein GAN(WGAN) (Arjovsky et al., 2017), WGAN with gradient penalty (WGAN-GP) (Gulrajani et al., 2017), and relativistic average GAN (RaGAN) (Jolicoeur-Martineau, 2019) are all optimized from the perspective of the loss function, which alleviated the instability of training to some extent. Since the image enhancement task can be considered an image-to-image translation task and the GANs can learn the mapping relationships between low- and high-quality domains; thus, GANs can be used for image enhancement.

2.2. Image enhancement

Image enhancement is the adjustment and optimization of attributes such as contrast, brightness, tone, and details of an image through several image processing techniques, enhancing the overall visual quality of the image.

The recent image enhancement tasks can be broadly divided into low-light image enhancement and image retouching. Previous studies have provided excellent ideas on image enhancement by using traditional methods, such as histogram equalization-based (Pizer et al., 1987) and Retinex-based methods (Li et al., 2018),(Guo et al., 2017). Histogram equalization and its variants are suitable for image brightness correction and contrast enhancement, which can be further divided into global histogram equalization (GHE) (Thomas et al., 2011) and local histogram equalization (LHE) (Lee et al., 2013). The former achieves overall image enhancement through a global mapping function; thus, it fails to focus on the local details of the images, leading to weaknesses such as local over-enhancement. The latter dynamically adjusts the histogram to solve the issue of local over-enhancement at the cost of its huge computational consumption. In addition, Retinex-based methods model image enhancement as an illumination estimation issue. Such approaches are based on the theory of Retinex, which assumes that all images can be represented as the product of the illumination and reflection components. Ma et al. in Ma et al. (2022) proposed a self-calibrated illumination (SCI) model with progressive optimization and parameter sharing based on Retinex theory, which corrects the inputs at each stage step-by-step and performs well on downstream tasks. Liang et al. in Liang et al. (2022) used an untrained neural network prior to enhancing low-light images with noise, which incorporates neural network-based co-optimization of Retinex decomposition and illumination adjustment, as well as illumination-guided self-supervised denoising. However, the estimation of the illumination component is very uncertain, and its inaccuracy will directly make the enhancement unnatural, lacking details and color distortion.

Convolutional neural networks (CNNs) have shown strong feature abstraction capability and are successfully applied in various computer vision tasks, including image enhancement. Chen et al. in Chen et al. (2018a) designed a fully convolutional network to enhance the raw sensor data directly, but it still suffers from the issue of color bias, and noise is evident. In addition, a combination of deep learning and traditional methods has also increased attention. Wei et al. in Wei et al. (2018) designed a deep Retinex-Net, which effectively combines the deep model with the Retinex model in an end-to-end framework for the feature decomposition and illumination adjustment. Ren et al. in Ren et al. (2019) introduced a hybrid network consisting of a content stream composed of an encoder–decoder structure and an edge stream composed of a spatially variant recurrent neural network (RNN), which learns global information without neglecting details. Cotogni et al. in Cotogni and Cusano (2023) proposed the first work to combine tree search theory with deep reinforcement learning for image enhancement. Guo et al. in Guo et al. (2020) treated image enhancement as a curve estimation problem, and such curves make pixel-level adjustments to the image over the dynamic range of the input without paired datasets. Jiang et al. in Jiang et al. (2021) designed a global–local discriminator and a self-regularized perceptual loss based on the GAN while incorporating prior knowledge. Sun et al. in Sun et al. (2021) designed a modulation code generator by formulating this ill-posed problem as a modulation code learning task. Moreover, the above methods are biased toward dealing with low-light images. Social media is rapidly developing nowadays, and more and more images appearing on the Internet require comprehensive adjustment of brightness, color, saturation, etc. Therefore, most of them cannot fully achieve the requirements of image retouching.

Bychkovsky et al. in Bychkovsky et al. (2011) provided a highly influential dataset named MIT-Adobe FiveK, and image retouching has attracted public attention. Ignatov et al. in Ignatov et al. (2017) also provided a dataset consisting of shots from phones and DSLRs showing the visual differences between devices. An end-to-end architecture was provided to implement feature mapping to improve visual perception. Ignatov et al. in Ignatov et al. (2018) proposed a weakly supervised image enhancement, which uses a CNN-GAN structure to learn mapping relationships. Chen et al. in Chen et al. (2018b) proposed a two-way GAN to improve the wasserstein GAN with an adaptive weighting scheme. He et al. in He et al. (2020) introduced an extremely light-weight framework based on pixel-level and space-level, but with extremely obvious modifications that may not fit with the popular aesthetic. Zamir et al. in Zamir et al. (2020) cared more about high-resolution spatial representations which can generate fine images. Jiang et al. in Jiang et al. (2020) proposed a coarse-to-fine structure from the translation perspective. For the same purpose, Liang et al. in Liang et al. (2021) incorporated the decomposition and reconstruction of the laplacian pyramid, which can reduce the computational cost and be used for high-resolution images. Ni et al. in Ni et al. (2020) extracted global statistics information to optimize the enhancement process, which is one of the few approaches that consider human aesthetics and perception. The work of Song et al. in Song et al. (2021) differed from other studies in that it focused on the flexible transformation of multiple styles. Based on our observations, it is necessary to use unpaired datasets for the training process in the task of image retouching, because the pairs are lacking in natural-world scenarios. Besides, the issues of noise, color bias, and lack of details have always been to be improved in the task of image retouching.

2.3. Laplacian pyramid

Laplacian pyramid (LP) (Burt and Adelson, 1983) is a classical image processing method that embodies multi-scale representation. Its primary thought is to reconstruct the undersampled image of the higher layers through the lower layers of the image pyramid, thus providing the maximum restoration of the image. Therefore, it can be regarded as an image decomposed into low-frequency signals and high-frequency details at different scales. It can be noticed that the focus of high-frequency information at various scales is different. As illustrated in Fig. 2, (b)–(e) are the high-frequency components of a set of LP,

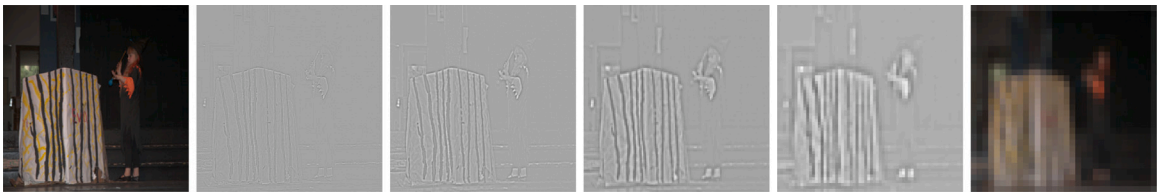


Fig. 2. A set of laplacian pyramid of an input image. (a) is the input image, (b)–(e) show the high-frequency component with a different scale, and (f) represents the low-frequency component. In fact, the size decreases gradually from (b) to (f). Here it is possible to visualize the differences between the detailed high-frequency information of the various layers preserved in the laplacian pyramid.

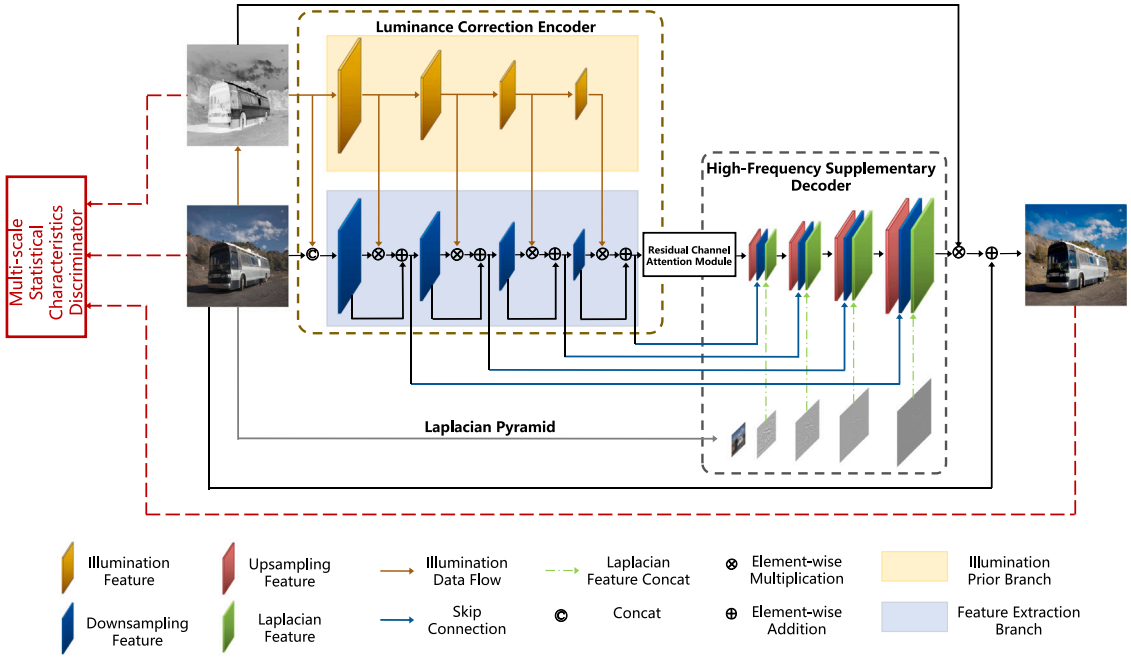


Fig. 3. The architecture of our proposed CAE-GAN. The luminance correction encoder (LCE), residual channel attention module (RCAM), and high-frequency supplementary decoder (HFSD) are elaborated in our generator. The luminance map stacks are fused with content features in the encoding process, and the appropriate brightness can be represented by the fused features. And the LP is gradually integrated at the decoder stage and serves to complement the details.

and (f) represents the low-frequency component. The low-frequency component reflects the underlying texture and color information, while the high-frequency component focuses on finer details that can be seen as the edge details. And the higher of layers, the finer the feedback details. And except for texture details, the light intensity of most pixels is close to 0. Due to the straightforward and excellent capability of image interpretation, the Laplacian pyramid is widely used in multiple fields. Due to the straightforward and excellent capability of image interpretation, the laplacian pyramid is widely used in image generation, super-resolution (Lai et al., 2017), semantic segmentation (Ghiasi and Fowlkes, 2016), image translation (Liang et al., 2021), medical image analysis (Han et al., 2022; Yin et al., 2023; Jian et al., 2023a,b), etc. And it can be noticed that both the low and high-frequency information can be effectively used for fine details reconstruction in image enhancement.

3. The proposed framework of CAE-GAN

This section presents detailed descriptions of our proposed CAE-GAN framework, which includes a core-attributes enhanced generator, a multi-scale statistical characteristics discriminator, and their various modules. In addition, an upgrade adversarial loss for this new discriminator structure is proposed to control global style, together with the perceptual loss and tone fidelity loss for the constraint.

3.1. Core-attributes enhanced generator

The core-attributes enhanced generator consists of a luminance correction encoder (LCE), a residual channel attention module (RCAM), and a high-frequency supplementary decoder (HFSD). The LCE is designed to adjust the brightness dynamically based on local differences to enhance the illumination. The RCAM captures the importance of channels while achieving enhancement, aiming to improve the capability of inter-channel mapping characterization. In the image reconstruction process, the HFSD complements the high-frequency details lost during the encoding process. Our main objective is to learn the mapping of low-quality domains to high-quality domains. The specific descriptions of our generator are shown in Fig. 3. An input sample from the low-quality domain can be represented as x. The image enhancement process can be represented as follows:

$$x' = G(x), \tag{1}$$

where G is the generator which is based on one-path GAN, and x' is the corresponding enhanced image of the high-quality domain.

3.1.1. Luminance correction encoder

The visual impact of lighting on an image is significant, and both over-dark and over-bright will affect the presentation of the content. In addition, the local brightness and darkness are different for each specific image. As shown in Fig. 4, if the brightness of the whole

(a) Input (b) Visualization of x_1^{lum}

Fig. 4. Here shows the input images and their corresponding x_1^{lum}. The brightness of the image varies from position to position, so the degree to which it needs to be adjusted is also different.

image is improved, the light source in the picture will be exposed. Thus, it is not reasonable to feed the whole image directly into the generator without any operation, because it lacks the specificity of the different parts. Therefore, the motivation of the luminance correction encoder is to extract the pixel-by-pixel luminance map of each image as prior knowledge to supervise the enhancement process, so that the network can dynamically adjust the brightness of each different image to achieve a high-quality illuminance enhancement effect.

As shown in Fig. 3, the LCE consists of two branches, the feature extraction branch (FEB) and the illumination prior branch (IPB). FEB extracts the input content features by stepwise downsampling and incorporates the luminance information from IPB. First, an RGB color image is converted into gray space to reflect the distribution of overall and local brightness levels, which is represented by different saturations of black for each image point. The conversion of RGB values and grayscale is achieved by (2)–(3) (Sector, 1995), which reflects the conversion of the human eye's perception of color to luminance. This process can be described as follows:

$$\left[\hat{x^R}, \hat{x^G}, \hat{x^B}\right] = \left[\frac{x^R}{255}, \frac{x^G}{255}, \frac{x^B}{255}\right], \tag{2}$$

where $\left[x^R, x^G, x^B\right]$ indicates the color signal of x in the RGB color space, $\left[\hat{x^R}, \hat{x^G}, \hat{x^B}\right]$ represents the normalized result. Thus, the grayscale image is reflected as:

$$x^{gray} = \hat{x^R} \times 0.299 + \hat{x^G} \times 0.587 + \hat{x^B} \times 0.114, \tag{3}$$

where x^{gray} denotes the grayscale image, which also reflects the brightness strength of input x. Therefore, the degree of brightness and darkness is transformed into the degree that needs to be noticed by subtraction as follows:

$$x_1^{lum} = 1 - x^{gray}, \tag{4}$$

where x_1^{lum} means the weight of the luminance to be improved, i.e., the darker pixels will get a substantial improvement in brightness. In the above way, x_1^{lum} is continuously downsampled to obtain a set of luminance weight stacks X^{lum} as the results of the IPB:

$$x_{i+1}^{lum} = Down_{lum_i}\left(x_i^{lum}\right), \tag{5}$$

$$X^{lum} = \{x_1^{lum}, x_{i+1}^{lum}\}, \tag{6}$$

where $i = 1, 2, 3$. $Down_{lum_i}(\cdot)$ stands for down-sampling operation, and X^{lum} contains four sets of luminance weights in different sizes, which are $1 \times 256 \times 256$, $1 \times 128 \times 128$, $1 \times 64 \times 64$ and $1 \times 32 \times 32$. Before extracting the content features, x and x_1^{lum} are concatenated and fed to the feature extraction branch. Then, X^{lum} is fully fused with the content features.

$$\begin{cases} F_i = E_i \cdot x_i^{lum} + E_i, \\ E_{i+1} = Down_{con_i}\left(F_i\right), \end{cases} \tag{7}$$

where $i = 1, 2, 3$, E_i means the extracted content features, the sizes from E_1 to E_4 are $32 \times 256 \times 256$, $64 \times 128 \times 128$, $128 \times 64 \times 64$ and $256 \times 32 \times 32$ respectively, and the first $E_1 = Concat\_conv(x, x_1^{lum})$, where $Concat\_conv(\cdot)$ stands for concatenation along the channel dimension and convolution operation. And F_i represents the fusion feature of i_{th} layer, while it will be used as input to get the content features of the next layer by down-sampling operation $Down_{con_i}(\cdot)$. "$\cdot$" indicates the element-wise multiplication.

3.1.2. Residual Channel Attention Module (RCAM)

The importance of the channels affects the quality of the image reconstruction, and useless channels do not help or even play a negative role in training. Therefore, the RCAM is designed to obtain global information, apply different attention to various regions of the global feature map, and calibrate the weights of channels adaptively. Inspired by Jie et al. (2017), Woo et al. (2018), two inter-channel dependencies are computed by the compressions of global max pooling (GMP) and global average pooling (GAP), and a weight vector of size $256 \times 1 \times 1$ is estimated. Then, the vector can be expanded in width and height to the size of $256 \times 32 \times 32$. The feature maps F_4 and the corresponding weights are multiplied to obtain the calibrated feature maps. This process can be described as follows:

$$RCAM\left(F_4\right) = Expand\left(fc\left(pool\left(F_4\right)\right)\right) \cdot F_4, \tag{8}$$

where F_4 is the feature maps extracted by LCE, and $pool(\cdot)$ means the combination of GMP and GAP. $fc(\cdot)$ is a fully connected layer, and $Expand(\cdot)$ stands for expanded operation. The specific structure is shown in Fig. 5.

3.1.3. High-frequency Supplementary Decoder (HFSD)

The human eye is very sensitive to high-frequency information, and its variation and absence can limit the acquired image information. Plenty of fine details were usually lost during the reconstruction, and the loss of high-frequency information in it resulted in a decline in image clarity. In the proposed HFSD, LP is utilized to enhance boundary contour details in a simple, straightforward, and effective way. For the input $x = I_1$, we decompose it into a set of LP components:

$$\begin{cases} lap_n = I_n - upsample(I_{n+1}), \\ I_{n+1} = Gaussian(I_n) \end{cases} \tag{9}$$

where $n = 1, 2, 3, 4$, $Gaussian(\cdot)$ is a weighted average operation of neighboring pixels based on a fixed Gaussian kernel, and $upsample(\cdot)$ means an up-sampling operation. In the HFSD, the LP has 5 levels. The first 4 layers are high-frequency residuals of different resolutions noted as $Lap = \{lap_1, lap_2, lap_3, lap_4\}$, and the last layer is a low-resolution image. Multiple features are concatenated together and expressed as:

$$\begin{cases} U_n = Up_{n-1}\left(D_{n-1}\right), \\ D_n = Conv_n\left(Concat\left(U_n, F_{5-n}, lap_n\right)\right), \end{cases} \tag{10}$$

where $n = 1, 2, 3, 4$, U_n means the up-sampling features, and the first $U_1 = RCAM\left(F_4\right)$. And D_n represents the fusion of three features, including the up-sampling features, the fusions of the illumination

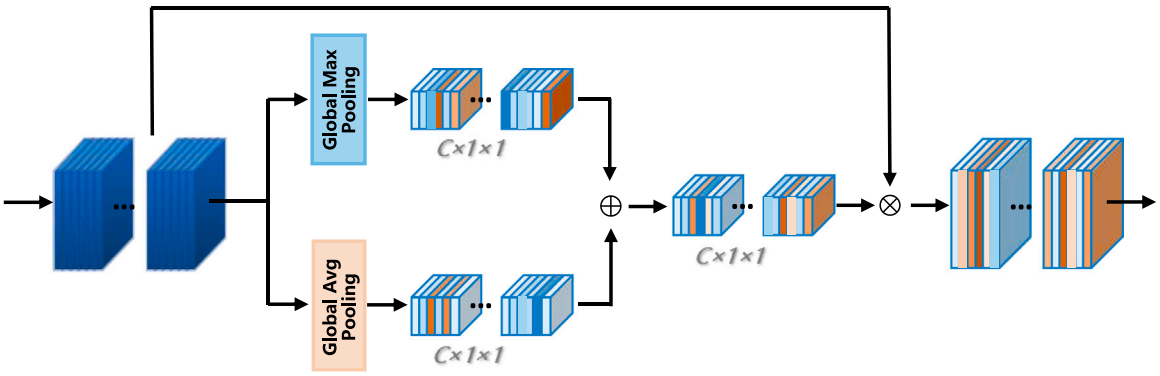


Fig. 5. The details of residual channel attention module (RCAM). It assigns weights according to the importance of the channel.

prior, and the high-frequency residuals. And $Conv_n(\cdot)$ is a descending operation on the channel dimension, $Concat(\cdot)$ stands for concatenation along the channel dimension. The difference compared to other methods is that additional shape transformations are not operated on the LP. This is because we observe that the high-frequency contour details hardly change before and after the image enhancement. Therefore, the details lost during encoding are added directly to the decoder, and other shape transformation operations are considered redundant. Moreover, skip connections (Ronneberger et al., 2015) preserve the structure information with F_{5-n}. As a consequence, we can obtain enhanced images that retain rich content.

3.2. Multi-scale statistical characteristics discriminator

The composition of an image is complex, especially since each local area has its own characteristics, such as local overexposure and darkness. To ensure that the images do not only conform to the target domain distribution as a whole, but also have a clear representation at the patch level, our discriminator is improving at PatchGAN (Isola et al., 2017). This structure enables full consideration of global and local image fidelity. The inputs to the discriminator are the unenhanced image x, the corresponding enhanced image x', or the randomly sampled target domain image y. For each input, it is discriminated at different layers, equivalent to focusing on patches of various sizes. The intermediate branch in Fig. 6 preserves PatchGAN for the purpose of multi-scale discrimination. In addition, we have followed the color attention module (CAM) designed in Liu et al. (2022), which can correct the global color tone through a self-attention mechanism (Zhang et al., 2019). This is reflected in the upper part of the branch in Fig. 6.

3.2.1. Multi-scale Statistical Characteristics Distinction Branch (MSCDB)

In this branch, we take two main points into consideration, including the color style and the full use of high and low-level information. From the perspective of style transfer, statistical features can be used to indicate the style of an image. A content image can be matched to a style image by transforming global statistical features. Therefore, MSCDB distinguishes domains in terms of global style and color tone. In addition, multi-scale statistical characteristics are conducive to improving stability at low-level, meanwhile, focusing on semantic discriminatory capabilities at high-level (Shao and Zhang, 2021). Therefore, a multi-scale statistical characteristics distinction branch (MSCDB) is designed to take advantage of the statistical characteristics of global and patch.

The MSCDB is shown in the bottom branch of Fig. 6. First, the feature a from different scales is fed into the adaptive modification module to be adapted for other dimensions, and a can be derived from x, x', and y. Next, the vectors of channel-wise statistical characteristics are calculated, specifically the mean and maximum values. Then, a final scalar result can be obtained by an MLP for each vector. This process can be abstracted as follows:

$$MSCDB\left(a_m\right) = MLP\left(SC\left(AM\left(a_m\right)\right)\right), \tag{11}$$

where $m = 1, 3, 5$ denotes different layers, $AM(\cdot)$ means the operation of adaptive modification consisting of convolutional layers, and $SC(\cdot)$ represents the calculations of statistical characteristics.

In summary, our discriminator incorporates three aspects: (1) classical multi-scale discrimination, (2) focused consideration of color tones, and (3) consideration of multi-scale statistical characteristics.

3.3. Loss function

In this subsection, our total goal is to minimize the following loss function, consisting of upgrade adversarial loss, perceptual loss, and tone fidelity loss:

$$L^G = \lambda_1 L^G_{adv} + \lambda_2 L_{per} + \lambda_3 L_{ton}, \tag{12}$$

where λ_1, λ_2, and λ_3 are the weights that balance the individual loss terms. Here we introduce each one of them.

3.3.1. Upgrade adversarial loss

The unenhanced image x, the corresponding enhanced image x', and the randomly sampled target domain image y are fed into our discriminator. We want to be able to obtain x' that matches the distribution where y is located. In addition, raHingeGAN (Jolicoeur-Martineau, 2019) is adopted instead of vanilla GAN (Goodfellow et al., 2014) because of its improved training stability. The objective of the discriminator is to recognize x and x' as fake whenever possible. This is expressed as (13):

$$\begin{aligned} L^D_1 =& \mathbb{E}_{y\sim P_y}\left[\max\left(0, 1-\left(D(y)-\mathbb{E}_{x\sim P_x}D(x)\right)\right)\right] \\ &+ \mathbb{E}_{x\sim P_x}\left[\max\left(0, 1+\left(D(x)-\mathbb{E}_{y\sim P_y}D(y)\right)\right)\right] \\ &+ \mathbb{E}_{y\sim P_y}\left[\max\left(0, 1-\left(D(y)-\mathbb{E}_{x\sim P_x}D(G(x))\right)\right)\right] \\ &+ \mathbb{E}_{x\sim P_x}\left[\max\left(0, 1+\left(D(G(x))-\mathbb{E}_{y\sim P_y}D(y)\right)\right)\right], \end{aligned} \tag{13}$$

where $G(\cdot)$ and $D(\cdot)$ correspond to the generator and discriminator, respectively, so $G(x) = x'$. In addition to L^D_1, we compute GAN loss L^D_2 on multiple statistical characteristics at multi-scales as (14):

$$\begin{aligned} L^D_2 =& \mathbb{E}_{y\sim P_y}\left[\max\left(0, 1-\left(D'(y)-\mathbb{E}_{x\sim P_x}D'(x)\right)\right)\right] \\ &+ \mathbb{E}_{x\sim P_x}\left[\max\left(0, 1+\left(D'(x)-\mathbb{E}_{y\sim P_y}D'(y)\right)\right)\right] \\ &+ \mathbb{E}_{y\sim P_y}\left[\max\left(0, 1-\left(D'(y)-\mathbb{E}_{x\sim P_x}D'(G(x))\right)\right)\right] \\ &+ \mathbb{E}_{x\sim P_x}\left[\max\left(0, 1+\left(D'(G(x))-\mathbb{E}_{y\sim P_y}D'(y)\right)\right)\right], \end{aligned} \tag{14}$$

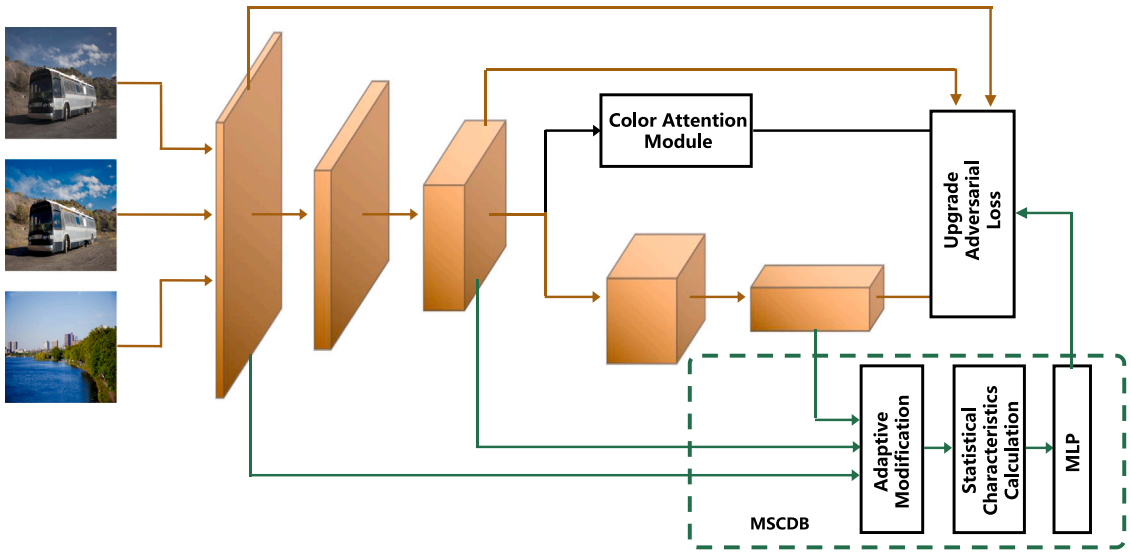


Fig. 6. The details of multi-scale statistical characteristics discriminator containing the MSCDB. It takes into account the multi-scale and multi-perspective discrimination of images. Specifically, we try the mean and maximum values.

where $D(\cdot)$ represents the traditional discrimination of features in Fig. 6, while $D'(\cdot)$ represents the discrimination of MSCDB branch. Therefore, the upgrade adversarial loss of the discriminator is:

$$L^D = L_1^D + \lambda L_2^D. \tag{15}$$

Likewise, the upgrade adversarial loss of the generator is also composed of two parts which can be calculated as (16), (17), and (18):

$$\begin{aligned} L_1^G =& \mathbb{E}_{x\sim P_x}\left[\max\left(0, 1-\left(D(G(x)) - \mathbb{E}_{y\sim P_y} D(y)\right)\right)\right] \\ &+\mathbb{E}_{y\sim P_y}\left[\max\left(0, 1+\left(D(y) - \mathbb{E}_{x\sim P_x} D(G(x))\right)\right)\right], \end{aligned} \tag{16}$$

$$\begin{aligned} L_2^G =& \mathbb{E}_{x\sim P_x}\left[\max\left(0, 1-\left(D'(G(x)) - \mathbb{E}_{y\sim P_y} D'(y)\right)\right)\right] \\ &+\mathbb{E}_{y\sim P_y}\left[\max\left(0, 1+\left(D'(y) - \mathbb{E}_{x\sim P_x} D'(G(x))\right)\right)\right], \end{aligned} \tag{17}$$

$$L_{adv}^G = L_1^G + \lambda L_2^G. \tag{18}$$

3.3.2. Perceptual loss

The semantics before and after enhancement should be consistent. The pixel-based distance is unreliable because it will encourage the output to be the same as the input, resulting in enhancement that will not appear in the result. It is contrary to the change of global style, so we add feature-based distance loss (Johnson et al., 2016) as (19):

$$L_{per} = \sum_{j=1}^{J} \mathbb{E}_{x\sim P_x}\left[\left\|\varphi_j(x) - \varphi_j(G(x))\right\|_2\right], \tag{19}$$

where $\varphi_j(\cdot)$ means the features extracted by the j_{th} layer of the pre-trained VGG-19. In our work, five layers *Relu*1_1, *Relu*2_1, *Relu*3_1, *Relu*4_1, *Relu*5_1 are adopted. Perceptual loss prevents the generator from over-emphasizing pixel consistency while disregarding image color enhancement effects, and preserving the content and structure of the original image.

3.3.3. Tone fidelity loss

For our generator, the expectation is that it can transform x into x' that obeys the high-quality domain distribution. To improve the stability of the generator, we randomly sample the image in the high-quality domain, denoted as y, and feed it into the generator. Since a low-quality to high-quality mapping is learned, the ideal result is that the generated y' remains almost pixel-level consistent with y. In summary, we consider that the relationship between x and x' is built on features, then y and y' are encouraged to establish a pixel-wise relationship. Therefore, we propose a tone fidelity loss as (20):

$$L_{ton} = \mathbb{E}_{y\sim P_y}\left\|y - G(y)\right\|_1. \tag{20}$$

where $G(y) = y'$ is the output of generator. L1 constraint is used to constrain the pixel differences effectively.

It is worth noting that since the high-quality image no longer requires qualitative improvements, the y does not require further luminance improvement when fed to the generator as input. However, some details are lost during the encoding process and additional high-frequency information is needed for it in the decoder. In fact, we set y to not go through the IPB, and the rest is kept in line with the process of x.

4. Experiments and setup

In this part, two well-known datasets are used for the performance evaluation. The proposed CAE-GAN is compared with a variety of classical and state-of-the-art methods, including both objective metrics and subjective visualization. In addition, adequate ablation experiments are also performed to estimate the effectiveness of the individual modules quantitatively and qualitatively.

4.1. Datasets

All experiments are trained and tested on the MIT-Adobe FiveK (Bychkovsky et al., 2011) and the DSLR photo enhancement dataset (DPED) (Ignatov et al., 2017).

- **MIT-Adobe FiveK:** The MIT-Adobe FiveK dataset contains 5000 low-quality images and the retouched counterparts from five professional photographers. The photographer C with the highest user recognition is chosen as labels in our work. It is randomly divided into three parts: (1) 2250 low-quality images and unpaired 2250 high-quality images for training; (2) 100 images for validation; (3) 400 images for testing. Although this dataset has labels, we did not use paired images during training, i.e., the 2250 low-quality domain images and the 2250 high-quality domain images do not overlap.
- **DPED:** The DPED consists of images captured by three smartphones and a Canon DSLR. In our training, we used the images captured by iPhone 3GS as low-quality images, and the

Table 1
Quantitative comparison on MIT-Adobe FiveK dataset. The evaluation metrics are PSNR, SSIM, VIF, FSIM, and LPIPS, respectively. The higher the first four metrics achieved, the more the quality of images matches the expectation, and the last metric vice versa. The top two are highlighted in red and blue colors, respectively.

Method	Metric				
	PSNR (↑)	SSIM (↑)	VIF (↑)	FSIM (↑)	LPIPS (↓)
Input	18.131	0.810	0.586	0.908	0.140
CycleGAN (Zhu et al., 2017)	21.708	0.815	0.584	0.950	0.139
EnlightenGAN (Jiang et al., 2021)	20.142	0.867	0.751	0.961	0.090
UEGAN (Ni et al., 2020)	23.936	0.890	0.755	0.970	0.070
CSRNet (He et al., 2020)	17.268	0.863	0.755	0.944	0.168
MIRNet (Zamir et al., 2020)	16.601	0.841	0.665	0.944	0.119
TSIT (Jiang et al., 2020)	18.306	0.843	0.479	0.899	0.168
Zero-DCE (Guo et al., 2020)	12.509	0.727	0.653	0.892	0.228
LPTN (Liang et al., 2021)	22.622	0.896	0.614	0.943	0.140
ImageEnhance_cGAN (Sun et al., 2021)	13.327	0.670	0.622	0.899	0.273
StarEnhancer (Song et al., 2021)	20.873	0.897	0.825	0.947	0.077
BNCAGAN (Liu et al., 2022)	24.332	0.910	0.760	0.971	0.068
SCI (Ma et al., 2022)	16.362	0.843	0.754	0.960	0.138
Untrained NN priors (Liang et al., 2022)	14.730	0.693	0.416	0.890	0.253
CAE-GAN(Ours)	24.979	0.914	0.753	0.973	0.069

Table 2
Quantitative comparison on DPED. The evaluation metrics are NIMA, IL-NIQE, and BRISQUE, respectively. The higher the first metric achieved, the more the quality of images matches the expectation, and the last two metrics vice versa. The top two are highlighted in red and blue colors, respectively.

Method	Metric		
	NIMA (↑)	IL-NIQE (↓)	BRISQUE (↓)
iPhone	4.058	23.707	17.675
CycleGAN (Zhu et al., 2017)	4.615	21.628	9.371
UEGAN (Jiang et al., 2021)	4.013	23.783	13.608
TSIT (Ni et al., 2020)	3.037	30.907	27.806
Zero-DCE (He et al., 2020)	4.240	22.486	15.305
LPTN (Jiang et al., 2020)	4.829	21.595	13.282
ImageEnhance_cGAN (Sun et al., 2021)	4.116	23.363	17.027
BNCAGAN (Liu et al., 2022)	4.658	21.429	8.949
CAE-GAN(Ours)	4.700	21.086	8.109

DSLR images were used as high-quality images. And each randomly selected 2500 images as the training set. The images taken by iPhone are generally poor in clarity and low in saturation, with yellowish picture quality. In comparison, the Canon DSLR captures the effect without the above problems.

4.2. Evaluation metrics

Since labels are available in the MIT-Adobe FiveK dataset and not in the DPED, we chose different evaluation metrics to measure them.

For the case with full reference, we used PSNR and SSIM (Wang et al., 2004), the most common and popular metrics for image quality assessment. However, many studies have concluded that the above two metrics do not exactly match the image quality profile observed by the human eye. Therefore, VIF (Sheikh and Bovik, 2006), FSIM (Zhang et al., 2011), and LPIPS (Zhang et al., 2018) are utilized to perform a comprehensive assessment. VIF quantifies the information which is extracted by the brain ideally from the reference image, and then measures the degree of distortion of this information. FSIM assigns various weights to the importance of pixels. And LPIPS is more in line with human perception than the above traditional methods. Multiple evaluation metrics are chosen to present the experimental results objectively from multiple perspectives.

For the DPED without standard labels, NIMA (Talebi and Milanfar, 2018), BRISQUE (Mittal et al., 2012), and IL-NIQE (Zhang et al., 2015) with no reference are chosen to evaluate. BRISQUE is the popular traditional evaluation algorithm that extracts the mean subtracted contrast normalized (MSCN) coefficients. And IL-NIQE is based on NIQE by integrating various aspects. Rating histograms can be generated by NIMA for any image, and direct comparisons between the same subject are made. It is close to the human sensory system which focuses on aesthetic quality.

4.3. The details of implementation

This work is implemented in Pytorch 1.7.1. We set the total epoch as 150, where the learning rate is 1e-4 for the first 75 epochs and decays linearly to 0 for the next 75 epochs. To ensure that the enhancement effect is as observed in Zhu et al. (2017), the batch size is set to 1. The λ_1, λ_2 and λ_3 in (12) are set as 0.1, 1.0, and 1.0 in the experiment, and the λ in (15) and (18) is set as 1.0. And the size of input and output is 512×512. Furthermore, in the preprocessing, cropped patches of size 256×256 are obtained randomly from the training set. Moreover, the inputs are randomly flipped and rotated for data augmentation.

4.4. Comparisons and ablation experiment

To show the excellent performance of our CAE-GAN, the results are compared with state-of-the-art methods as follows: CycleGAN (Zhu et al., 2017), EnlightenGAN (Jiang et al., 2021), UEGAN (Ni et al., 2020), CSRNet (He et al., 2020), MIRNet (Zamir et al., 2020), TSIT (Jiang et al., 2020), Zero-DCE (Guo et al., 2020), LPTN (Liang et al., 2021), ImageEnhance_cGAN (Sun et al., 2021), StarEnhancer (Song et al., 2021), BNCAGAN (Liu et al., 2022), SCI (Ma et al., 2022), and Untrained NN Priors (Liang et al., 2022). Here we will analyze the objective results and visual effects.

4.4.1. Quantitative comparisons and analyses

As shown in the Tables 1 and 2, the experimental results of each method on the MIT-Adobe FiveK dataset and DPED are demonstrated separately. Among all these compared methods, we use the open-source code of the original authors to train on our devices, except for a few methods where the models are publicly available.

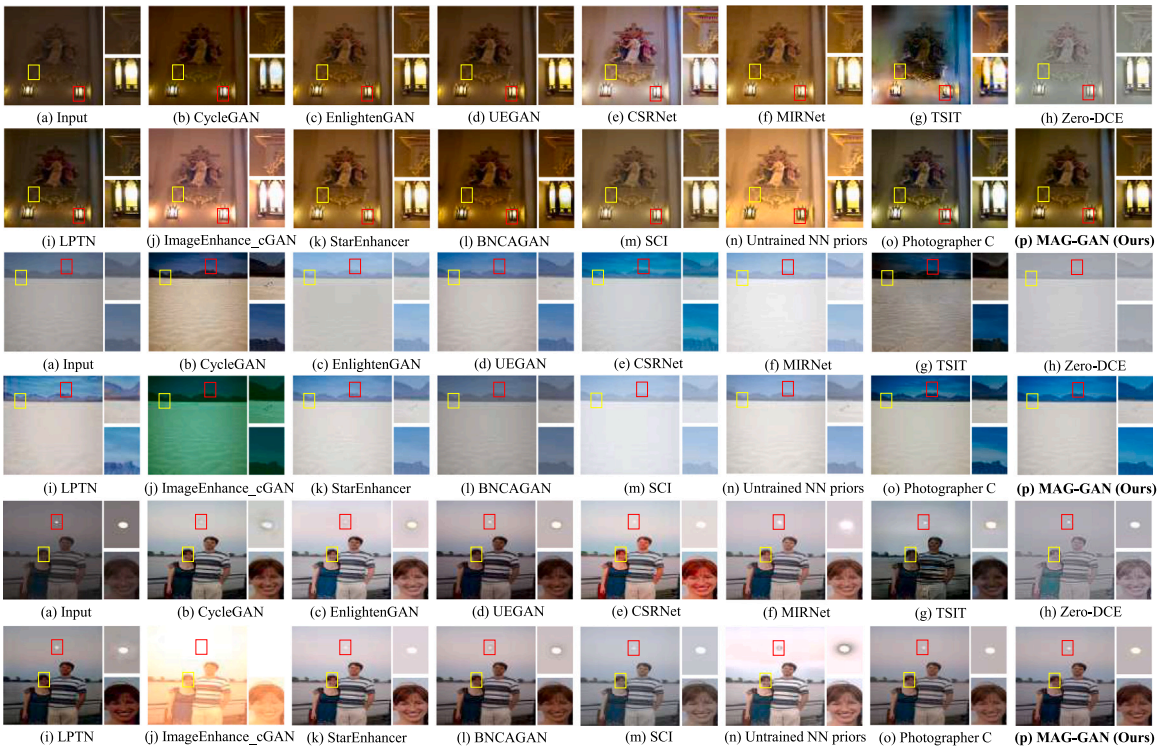


Fig. 7. The visual comparison with other state-of-the-art methods on MIT-Adobe FiveK dataset. Here we compare with CycleGAN (Zhu et al., 2017), EnlightenGAN (Jiang et al., 2021), UEGAN (Ni et al., 2020), CSRNet (He et al., 2020), MIRNet (Zamir et al., 2020), TSIT (Jiang et al., 2020), Zero-DCE (Guo et al., 2020), LPTN (Liang et al., 2021), ImageEnhance_cGAN (Sun et al., 2021), StarEnhancer (Song et al., 2021), BNCAGAN (Liu et al., 2022), SCI (Ma et al., 2022), and Untrained NN Priors (Liang et al., 2022), respectively. At the same time, the inputs and their modifications of photographer C are shown.

The goal of the experiment on the MIT-Adobe FiveK dataset is to get enhanced images having the style of photographer C. In the Table 1, our CAE-GAN is superior to other methods in terms of PSNR, SSIM, and FSIM, and takes second place on LPIPS. Compared to the input, we can see that enhancement performance significantly improves by 6.848db on PSNR, and improves by 0.104 on SSIM. It also increases by 0.167, 0.065, and 0.071 on VIF, FSIM, and LPIPS, respectively. This shows that CAE-GAN does have a noticeable effect on image quality.

The comparison experiments we selected can be divided into three categories. One is the image retouching task, which is the closest to our work. The result for UEGAN stands out among the multiple results, but is slightly lower than ours. CSRNet does not improve in the results, because it over-improved in contrast and deviated from our target style. The enhancement effect of MIRNet is very insignificant in the manifestation of the metrics. Because it is effective for denoising but lacks learning of styles, through attention and multi-scale residual blocks. And ImageEnhance_cGAN even plays a reverse role. StarEnhancer greatly improved in VIF, but the other metrics are performing average, indicating that it may perform well visually but deviates from photographer C's style. BNCAGAN is second only to CAE-GAN, which focuses on overall style improvement, while CAE-GAN is more specific for each attribute. The second is the low-light image enhancement task, which can be compared to our effect in brightness enhancement. The limitation of EnlightenGAN is that it can only increase the brightness, and Zero-DCE is similar. This type of method may only work for extremely dark images and will over-brighten images that are less dark, resulting in poor enhancement. The last one is the image-to-image translation (I2I) method that can be applied to multiple subtasks, including low-quality domain to high-quality domain transformation. CycleGAN, as a classical I2I algorithm, can only achieve enhancement effects in general and lacks detail. And TSIT has the same problem. And LPTN works better because it also considers the high-frequency information of the image. Still, the reconstruction from low-resolution images sacrifices the transformation effect accordingly, even though it reduces the computational effort. Untrained NN priors lack supervised information during the training process and therefore are not as effective as they could be.

The purpose of the experiment on DPED is to convert the quality taken by the iPhone to the quality achieved by the Canon DSLR. It is relatively challenging to learn the mapping from the iPhone domain to the DSLR domain because their differences are not significant. UEGAN does not perform well on DPED and has little to no improvement, and it may be insensitive to subtle domain differences. ImageEnhance_cGAN abstracts the enhancement process into a modulation code guided image generation task, where the quality of the modulation code is related to the stability of the final result. BNCAGAN works well in terms of metrics, but is slightly less effective than CAE-GAN overall. Algorithms like Zero-DCE that improve brightness tend to over-enhance the image. CycleGAN is more suitable for detailed tuning than transformations with widely varying domains, thus effective on this dataset. Conversely, TSIT is more suitable for I2I tasks with large deformations, so the lack of fine details leads to poor performance. LPTN performs well on NIMA, while the other two metrics are not satisfactory, indicating that it is superior in aesthetic quality. However, it may deviate from our purpose of learning the DSLR domain. In conclusion, our CAE-GAN is more balanced in each metric.

4.4.2. *Visual comparisons and analyses*

To further visualize the performance of our CAE-GAN, we compare the visual effects with other methods. As shown in Figs. 7 and 8, here are the test results on the MIT-Adobe FiveK dataset and DPED, respectively.

In Fig. 7, the input images are characterized by dark lighting, lack of detail, and low saturation. UEGAN has severe artifacts and is insensitive in luminance enhancement, which has a very strong impact on visual perception. CSRNet has excellent contrast and saturation but deviates from photographer C's style, which is not the representation we want. And MIRNet produces exposures on bright blocks resulting in

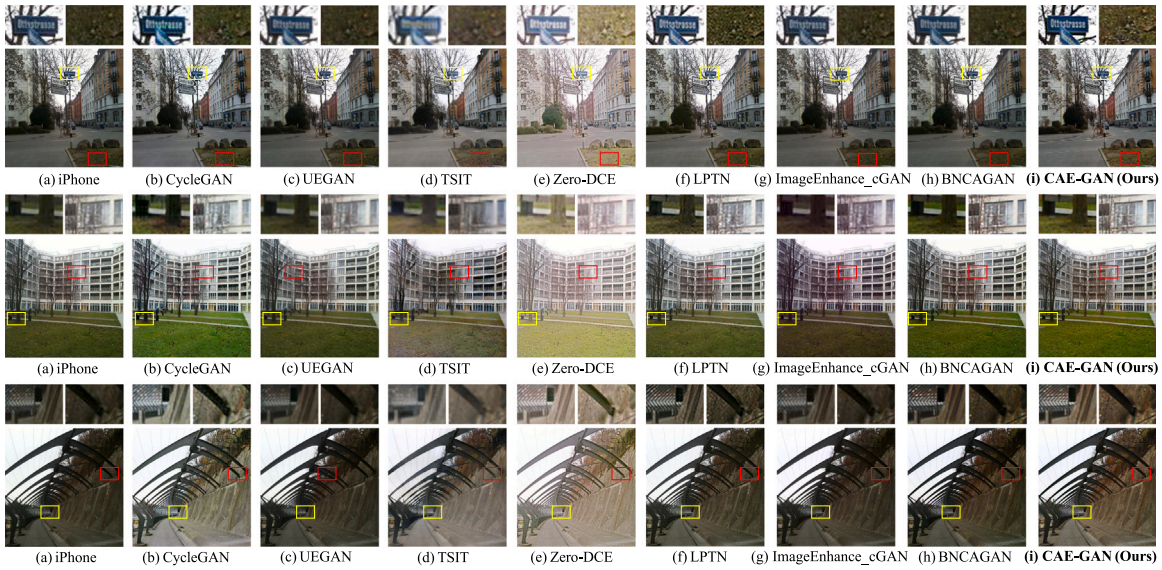


Fig. 8. The visual comparison with other state-of-the-art methods on DPED. Here we compare with CycleGAN (Zhu et al., 2017), UEGAN (Ni et al., 2020), TSIT (Jiang et al., 2020), Zero-DCE (Guo et al., 2020), LPTN (Liang et al., 2021), ImageEnhance_cGAN (Sun et al., 2021), and BNCAGAN (Liu et al., 2022), respectively.

Fig. 9. The visual performance for each part on the MIT-Adobe FiveK dataset. From (a) to (f) corresponds to the case where different modules are added. It is obvious that CAE-GAN has the most outstanding performance in brightness, tones, and details.

visual degradation. ImageEnhance_cGAN is the worst performer among all image retouching methods whose instability is reflected in the difficulty of capturing the target domain distribution. StarEnhancer works visually very satisfyingly, except for the fact that its color tone does not match what we expect. Moreover, BNCAGAN performs well in details and tones, but lacks in brightness and is insensitive to some hard samples. EnlightenGAN only improves on brightness but performs feebly on color. And Zero-DCE amplifies the above disadvantage, and it is not adaptive for different degrees of brightness and darkness. The outputs of CycleGAN are very rough because the cycle consistency loss is tightly constrained and has little stability. TSIT not only plays a minor role in detailed textures and tones but even destroys the original content and structural information. LPTN performs poorly in detail and has artifacts on the edge of the outline, with many uneven color blocks, also proving that this method is not well adapted to such a delicate task as image enhancement. SCI is missing in the color of the images, and Untrained NN priors produce severe artifacts. In contrast, CAE-GAN presents the most comprehensive and visually optimal results.

The comparisons on DPED are shown in Fig. 8, the images captured by iPhone are characterized by blur and yellowing. UEGAN has almost no effect of enhancement on this dataset, which shows the limitation of the scope to which this method is applicable. And ImageEnhance_cGAN is not only that but also a tone deviation occurs. BNCAGAN has progress to be made in terms of brightness compared to CAE-GAN, and demonstrates that our proposed LCE does have an advantage in terms of brightness enhancement. As always, Zero-DCE demonstrates high-intensity brightness enhancement. Besides, CycleGAN grasps the general direction of the conversion, but the results do not bear scrutiny due to the existence of artifacts. TSIT increases brightness at the cost of clarity, and LPTN gets clarity lacking rich colors. Overall, our CAE-GAN is the most balanced approach to performance, in all aspects of detail sharpness, color saturation, and brightness.

4.4.3. Experiment of ablation

In this subsubsection, we further demonstrate experimentally the contribution of each of the three components proposed in this work. In addition, we explore the layers selected by MSCDB in the discriminator as well as the statistical Characteristics to show their respective influence.

As shown in Table 3, here is the effect of IPB, LP, and MSCDB, and Fig. 9 shows the visual difference. IPB and LP each have a slight increase in the metrics, which have increased the PSNR by 0.589db and 0.604db respectively. And the combination of the two is a significant increase, which has improved 1.035db on PSNR. As shown in Table 3, MSCDB has a great improvement on VIF, increasing by 0.036. The results reach the highest with the correction of IPB, LP, and MSCDB,

Table 3
Quantitative comparison of our method with each module on MIT-Adobe FiveK dataset.

Method	Module			Metric				
	IPB	LP	MSCDB	PSNR (↑)	SSIM (↑)	VIF (↑)	FSIM (↑)	LPIPS (↓)
F1				23.569	0.897	0.737	0.967	0.082
F2	✓			24.158	0.883	0.746	0.970	0.075
F3		✓		24.173	0.905	0.761	0.971	0.074
F4	✓	✓		24.604	0.906	0.752	0.972	0.071
F5			✓	23.251	0.898	0.773	0.969	0.073
CAE-GAN	✓	✓	✓	24.979	0.914	0.753	0.973	0.069

Table 4
Quantitative comparison of our method with each layer and statistical characteristic in the discriminator on MIT-Adobe FiveK dataset.

Method	Layer			Mean	Max	Metric				
	lay_1	*lay_3*	*lay_5*			PSNR (↑)	SSIM (↑)	VIF (↑)	FSIM (↑)	LPIPS (↓)
F6	✓			✓	✓	24.448	0.902	0.759	0.972	0.071
F7	✓	✓		✓	✓	24.589	0.902	0.756	0.970	0.079
F8	✓	✓	✓	✓		24.616	0.904	0.757	0.972	0.071
F9	✓	✓	✓		✓	24.775	0.910	0.748	0.972	0.070
CAE-GAN	✓	✓	✓	✓	✓	24.979	0.914	0.753	0.973	0.069

Fig. 10. The visual performance for different layers and statistical characteristics in our discriminator on the MIT-Adobe FiveK dataset. (a), (b), and (e) correspond to the case where different layers are added. (c), (d), and (e) correspond to the case where different statistical characteristics are considered.

which have significant improvements in PSNR, SSIM, and LPIPS by 0.531db, 0.012, and 0.002. In Fig. 9, The F2 adds IPB to the F1, and the brightness of the image has indeed improved greatly, especially in contrast between light and dark. The F3 adds LP to the F1, improving the original unnatural edge contours and details. F4, with the help of IPB and LP, the overall effect is more beautiful, but there are still uneven blocks of color when you look carefully. Finally, with the role of MSCDB, the above problem can be solved. It is worth noting that using MSCDB alone in F5, the metrics will drop, probably because of the lack of control over the details, as shown in Fig. 9, there are many artifacts in the details around the boat. CAE-GAN is able to focus on brightness, details, and color tones at the same time.

To further explore the role of different layers and statistical characteristics in the MSCDB, features of various depths are taken out and fed into the MSCDB in a discriminator of five-layer. The results of the comparison are shown in Table 4 and Fig. 10. Comparing F6, F7, and CAE-GAN, as deeper features continue to be taken into account, the metrics are constantly being improved. In addition, while both the mean and maximum values alone have varying degrees of positive effect, the combination has the best performance and also shows that fusing them can be balanced.

5. Discussion

The proposed CAE-GAN can effectively improve brightness enhancement, image sharpness, and color correction for images in natural scenes. During the encoding process, the pixel-by-pixel brightness prior can obtained by calculating the low-quality image, and provide guidance for the network to dynamically adjust the brightness of each area according to the characteristics of different images. This ensures that our results do not have the problem of over-bright. Meanwhile, the image edge sharpness is improved by incorporating high-frequency information in the decoder. It can directly and simply restore the sharpness of the image, so it is a great improvement over other image enhancement methods that do not consider sharpness. Through the iterated training of discriminator and generator, we can get bright and clear images. In addition, the MSCDB can enhance the attention of the global style, which focuses on the tone information between the three color channels. Therefore, CAE-GAN realizes the comprehensive enhancement effect of multiple attributes at once, rather than a single attribute.

However, there are inevitable limitations due to its unsupervised training method. As shown in Fig. 11, the typical problem is that some of the images, while visually satisfying, are tonally inconsistent with the labels. This limitation is generally seen in cases where low-quality images span a wide range of tonal styles from the labels. In addition, the LCE significantly affects brightness enhancement, so it is a matter of concern whether it will affect some black objects in the image. As shown in Fig. 11, the negative impact of CAE-GAN on black objects when brightening images is small and almost negligible.

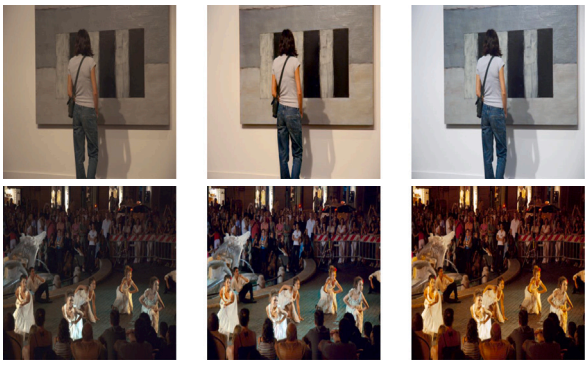


Fig. 11. Failure cases generated by our method compared with the labels.

6. Conclusion

This paper proposes a novel core-attributes enhanced generative adversarial network (CAE-GAN) for high-quality image enhancement without paired dataset. The CAE-GAN consists of a luminance correction encoder (LCE), a high-frequency supplementary decoder (HFSD), and a multi-scale statistical characteristics distinction branch (MSCDB). The effectiveness and superiority of our method are verified through extensive experiments. Specifically, the LCE effectively adjusts the brightness required for each location dynamically and adaptively, obtaining a high-quality illuminance effect. The HFSD timely supplements details to the edge contour during the training process, significantly improving the enhanced image's sharpness effect. In addition, the designed MSCDB effectively controls global style in the discriminator, considering the stability of low-level features and the strong discriminative capability of high-level features. Furthermore, the enhanced performance of CAE-GAN is competitive or even better compared to other excellent methods. Our current work is to improve image quality from the perspective of brightness, edge detail, and color style. Therefore, future work is to pay attention to more image attributes, such as saturation and contrast, and the challenge of considering so many factors entirely. Meanwhile, how to solve the problem of color deviation in images with large tonal gaps in unsupervised image enhancement methods is a challenging direction.

CRediT authorship contribution statement

Shan Liu: Conceptualization, Methodology, Software, Investigation, Writing – original draft. **Guoqiang Xiao:** Review & editing. **Michael S. Lew:** Review & editing. **Xinbo Gao:** Review & editing. **Song Wu:** Conceptualization, Methodology, Software, Writing – review & editing.

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.

Data availability

Data will be made available on request.

Acknowledgment

This work was supported by the Fundamental Research Funds for the Central Universities, China (SWU-KT22032).

References

Arjovsky, M., Chintala, S., Bottou, L., 2017. Wasserstein GAN.
Burt, P., Adelson, E., 1983. The Laplacian pyramid as a compact image code. IEEE Trans. Commun. 31 (4), 532–540.
Bychkovsky, Vladimir, Paris, Sylvain, Chan, Eric, Durand, Fredo, 2011. Learning photographic global tonal adjustment with a database of input / output image pairs. In: CVPR 2011. pp. 97–104.
Chen, Chen, Chen, Qifeng, Xu, Jia, Koltun, Vladlen, 2018a. Learning to see in the dark. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3291–3300.
Chen, Yu-Sheng, Wang, Yu-Ching, Kao, Man-Hsin, Chuang, Yung-Yu, 2018b. Deep photo enhancer: Unpaired learning for image enhancement from photographs with GANs. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6306–6314.
Cotogni, Marco, Cusano, Claudio, 2023. TreEnhance: A tree search method for low-light image enhancement. Pattern Recognit. 136, 109249.
Ghiasi, Golnaz, Fowlkes, Charless C., 2016. Laplacian pyramid reconstruction and refinement for semantic segmentation. In: Computer Vision – ECCV 2016. Vol. 9907, Springer, pp. 519–534.
Goodfellow, I.J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D, Ozair, S., Courville, A., Bengio, Y., 2014. Generative adversarial networks. Adv. Neural Inf. Process. Syst. 3, 2672–2680.
Gulrajani, Ishaan, Ahmed, Faruk, Arjovsky, Martín, Dumoulin, Vincent, Courville, Aaron C., 2017. Improved training of wasserstein GANs. In: Proceedings of the 30th International Conference on Neural Information Processing Systems. pp. 5767–5777.
Guo, Chunle, Li, Chongyi, Guo, Jichang, Loy, Chen Change, Hou, Junhui, Kwong, Sam, Cong, Runmin, 2020. Zero-reference deep curve estimation for low-light image enhancement. In: CVPR 2020. pp. 1777–1786.
Guo, Xiaojie, Li, Yu, Ling, Haibin, 2017. LIME: Low-light image enhancement via illumination map estimation. IEEE Trans. Image Process. 26 (2), 982–993.
Han, Zhimeng, Jian, Muwei, Wang, Gai-Ge, 2022. ConvUNeXt: An efficient convolution neural network for medical image segmentation. Knowl.-Based Syst. 253, 109512.
He, Jingwen, Liu, Yihao, Qiao, Yu, Dong, Chao, 2020. Conditional sequential modulation for efficient global image retouching. In: ECCV 2020. pp. 679–695.
Huang, Xun, Belongie, Serge, 2017. Arbitrary style transfer in real-time with adaptive instance normalization. In: 2017 IEEE International Conference on Computer Vision (ICCV). pp. 1510–1519. http://dx.doi.org/10.1109/ICCV.2017.167.
Ignatov, Andrey, Kobyshev, Nikolay, Timofte, Radu, Vanhoey, Kenneth, 2017. DSLR-quality photos on mobile devices with deep convolutional networks. In: ICCV 2017. pp. 3297–3305.
Ignatov, Andrey, Kobyshev, Nikolay, Timofte, Radu, Vanhoey, Kenneth, Gool, Luc Van, 2018. WESPE: Weakly supervised photo enhancer for digital cameras. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW).
Isola, Phillip, Zhu, Jun-Yan, Zhou, Tinghui, Efros, Alexei A., 2017. Image-to-image translation with conditional adversarial networks. In: CVPR 2017. pp. 5967–5976.
Jian, Muwei, Chen, Hongyu, Tao, Chen, Li, Xiaoguang, Wang, Gaige, 2023a. Triple-DRNet: A triple-cascade convolution neural network for diabetic retinopathy grading using fundus images. Comput. Biol. Med. 155, 106631.
Jian, Muwei, Wu, Ronghua, Chen, Hongyu, Fu, Lanqi, Yang, Chengdong, 2023b. Dual-branch-unet: A dual-branch convolutional neural network for medical image segmentation. CMES-Comput. Model. Eng. Sci. 137 (1).
Jiang, Yifan, Gong, Xinyu, Liu, Ding, Cheng, Yu, Fang, Chen, Shen, Xiaohui, Yang, Jianchao, Zhou, Pan, Wang, Zhangyang, 2021. Enlightengan: Deep light enhancement without paired supervision. IEEE Trans. Image Process. 30, 2340–2349.
Jiang, Liming, Zhang, Changxu, Huang, Mingyang, Liu, Chunxiao, Shi, Jianping, Loy, Chen Change, 2020. TSIT: A simple and versatile framework for image-to-image translation. In: ECCV 2020. pp. 206–222.
Jie, H., Li, S., Gang, S., Albanie, S., 2017. Squeeze-and-excitation networks. In: IEEE.
Johnson, Justin, Alahi, Alexandre, Fei-Fei, Li, 2016. Perceptual losses for real-time style transfer and super-resolution. In: ECCV 2016. pp. 694–711.
Jolicoeur-Martineau, Alexia, 2019. The relativistic discriminator: a key element missing from standard GAN. In: ICLR 2019.
Lai, Wei-Sheng, Huang, Jia-Bin, Ahuja, Narendra, Yang, Ming-Hsuan, 2017. Deep Laplacian pyramid networks for fast and accurate super-resolution. In: CVPR 2017. pp. 5835–5843.
Ledig, Christian, Theis, Lucas, Huszár, Ferenc, Caballero, Jose, Cunningham, Andrew, Acosta, Alejandro, Aitken, Andrew, Tejani, Alykhan, Totz, Johannes, Wang, Zehan, Shi, Wenzhe, 2017. Photo-realistic single image super-resolution using a generative adversarial network. In: CVPR 2017. pp. 105–114.
Lee, Chulwoo, Lee, Chul, Kim, Chang-Su, 2013. Contrast enhancement based on layered difference representation of 2D histograms. IEEE Trans. Image Process. 22 (12), 5372–5384.
Li, Mading, Liu, Jiaying, Yang, Wenhan, Sun, Xiaoyan, Guo, Zongming, 2018. Structure-revealing low-light image enhancement via robust retinex model. IEEE Trans. Image Process. 27 (6), 2828–2841.

Liang, Jinxiu, Xu, Yong, Quan, Yuhui, Shi, Boxin, Ji, Hui, 2022. Self-supervised low-light image enhancement using discrepant untrained network priors. IEEE Trans. Circuits Syst. Video Technol. Early Access.

Liang, Jie, Zeng, Hui, Zhang, Lei, 2021. High-resolution photorealistic image translation in real-time: A Laplacian pyramid translation network. In: CVPR 2021. pp. 9387–9395.

Liu, Shan, Xiao, Guoqiang, Xu, Xiaohui, Wu, Song, 2022. Bi-directional normalization and color attention-guided generative adversarial network for image enhancement. In: 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 2205–2209. http://dx.doi.org/10.1109/ICASSP43922.2022.9746840.

Luo, Pinjun, Xiao, Guoqiang, Gao, Xinbo, Wu, Song, 2023. LKD-net: Large kernel convolution network for single image dehazing. In: IEEE International Conference on Multimedia and Expo. pp. 1601–1606.

Ma, Long, Ma, Tengyu, Liu, Risheng, Fan, Xin, Luo, Zhongxuan, 2022. Toward fast, flexible, and robust low-light image enhancement. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5637–5646.

Mao, Xudong, Li, Qing, Xie, Haoran, Lau, Raymond Y.K., Wang, Zhen, Smolley, Stephen Paul, 2017. Least squares generative adversarial networks. In: 2017 IEEE International Conference on Computer Vision (ICCV). pp. 2813–2821.

Mittal, Anish, Moorthy, Anush Krishna, Bovik, Alan Conrad, 2012. No-reference image quality assessment in the spatial domain. IEEE Trans. Image Process. 21 (12), 4695–4708.

Ni, Zhangkai, Yang, Wenhan, Wang, Shiqi, Ma, Lin, Kwong, Sam, 2020. Towards unsupervised deep image enhancement with generative adversarial network. IEEE Trans. Image Process. 29, 9140–9151.

Pizer, Stephen M., Amburn, E. Philip, Austin, John D., Cromartie, Robert, Geselowitz, Ari, Greer, Trey, ter Haar Romeny, Bart, Zimmerman, John B., Zuiderveld, Karel, 1987. Adaptive histogram equalization and its variations. Comput. Vis. Graph. Image Process. 39 (3), 355–368.

Radford, Alec, Metz, Luke, Chintala, Soumith, 2015. Unsupervised representation learning with deep convolutional generative adversarial networks. Comput. Sci..

Ren, Wenqi, Liu, Sifei, Ma, Lin, Xu, Qianqian, Xu, Xiangyu, Cao, Xiaochun, Du, Junping, Yang, Ming-Hsuan, 2019. Low-light image enhancement via a deep hybrid network. IEEE Trans. Image Process. 28 (9), 4364–4375.

Ronneberger, Olaf, Fischer, Philipp, Brox, Thomas, 2015. U-Net: Convolutional networks for biomedical image segmentation. In: Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015. Springer International Publishing, Cham, pp. 234–241.

Sector, Itur, 1995. Studio encoding parameters of digital television for standard 4 : 3 and wide-screen 16 : 9 aspect ratios. International Telecommunication Union Radiocommunications Sector (ITU-R), BT.601-5.

Shao, X., Zhang, W., 2021. SPatchGAN: A statistical feature based discriminator for unsupervised image-to-image translation.

Sheikh, H.R., Bovik, A.C., 2006. Image information and visual quality. IEEE Trans. Image Process. 15 (2), 430–444.

Song, Yuda, Qian, Hui, Du, Xin, 2021. StarEnhancer: Learning real-time and style-aware image enhancement. In: ICCV 2021. pp. 4106–4115.

Sun, X., Li, M., He, T., Fan, L., 2021. Enhance image as you like with unpaired learning. In: IJCAI.

Talebi, Hossein, Milanfar, Peyman, 2018. NIMA: Neural image assessment. IEEE Trans. Image Process. 27 (8), 3998–4011.

Thomas, Gabriel, Flores-Tapia, Daniel, Pistorius, Stephen, 2011. Histogram specification: A fast and flexible method to process digital images. IEEE Trans. Instrum. Meas. 60 (5), 1565–1578.

Wang, Zhou, Bovik, A.C., Sheikh, H.R., Simoncelli, E.P., 2004. Image quality assessment: from error visibility to structural similarity. IEEE Trans. Image Process. 13 (4), 600–612.

Wei, C., Wang, W., Yang, W., Liu, J., 2018. Deep retinex decomposition for low-light enhancement. In: British Machine Vision Conference (BMVC).

Woo, Sanghyun, Park, Jongchan, Lee, Joon-Young, Kweon, In So, 2018. CBAM: Convolutional block attention module. In: ECCV 2018. pp. 3–19.

Xu, Xiaohui, Liu, Shan, Zhang, Nian, Xiao, Guoqiang, Wu, Song, 2022. Channel exchange and adversarial learning guided cross-modal person re-identification. Knowl.-Based Syst. 109883.

Yin, Yunchou, Han, Zhimeng, Jian, Muwei, Wang, Gai-Ge, Chen, Liyan, Wang, Rui, 2023. AMSUnet: A neural network using atrous multi-scale convolution for medical image segmentation. Comput. Biol. Med. 162, 107120, URL https://www.sciencedirect.com/science/article/pii/S0010482523005851.

Zamir, Syed Waqas, Arora, Aditya, Khan, Salman, Hayat, Munawar, Khan, Fahad Shahbaz, Yang, Ming-Hsuan, Shao, Ling, 2020. Learning enriched features for real image restoration and enhancement. In: ECCV 2020. pp. 492–511.

Zhang, Han, Goodfellow, Ian, Metaxas, Dimitris, Odena, Augustus, 2019. Self-attention generative adversarial networks. In: International Conference on Machine Learning. PMLR, pp. 7354–7363.

Zhang, Richard, Isola, Phillip, Efros, Alexei A., Shechtman, Eli, Wang, Oliver, 2018. The unreasonable effectiveness of deep features as a perceptual metric. In: CVPR 2018. pp. 586–595.

Zhang, Lin, Zhang, Lei, Bovik, Alan C., 2015. A feature-enriched completely blind image quality evaluator. IEEE Trans. Image Process. 24 (8), 2579–2591.

Zhang, Lin, Zhang, Lei, Mou, Xuanqin, Zhang, David, 2011. FSIM: A feature similarity index for image quality assessment. IEEE Trans. Image Process. 20 (8), 2378–2386.

Zhu, Jun-Yan, Park, Taesung, Isola, Phillip, Efros, Alexei A., 2017. Unpaired image-to-image translation using cycle-consistent adversarial networks. In: ICCV 2017. pp. 2242–2251.

Zou, X., Wu, S., Zhang, N., Bakker, E.M., 2022. Multi-label modality enhanced attention based self-supervised deep cross-modal hashing. Knowl.-Based Syst. 239, 107927.