IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received March 17, 2021, accepted March 23, 2021, date of publication March 24, 2021, date of current version April 2, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3068861

Joint Learning of Super-Resolution and Perceptual Image Enhancement for Single Image

YIFEI XU[1,4], **NUO ZHANG**[1], **LI LI**[2], **GENAN SANG**[2], **YUEWAN ZHANG**[1], **ZHENGYANG WANG**[1], **AND PINGPING WEI**[3]
[1]School of Software, Xi'an Jiaotong University, Xi'an 710054, China
[2]Alltuu Inc., Hangzhou 311100, China
[3]State Key Laboratory for Manufacturing Systems Engineering, Xi'an Jiaotong University, Xi'an 710054, China
[4]Huiyichen Inc., Nanchang 330038, China

Corresponding author: Yifei Xu (belonxu_1@xjtu.edu.cn)

This work was supported in part by the Young Scientists Fund of the National Natural Science Foundation of China under Grant 61802300, in part by the China Postdoctoral Science Foundation Funded Project under Grant 2018m643666, in part by the Xi'an Jiaotong University Basic Research Foundation for Young Teachers under Grant xjh012019043, and in part by the National Science and Technology Major Project under Grant 2019YFB2102501 and Grant 2019YFB2103005.

ABSTRACT Super resolution (SR) and Perceptual Image Enhancement (PIE) are gaining more and more interests in digital image processing and have been studied independently in the past decades. Although plenty of state-of-the-art researches have demonstrated great improvement in SR problem, they neglect practical requirements in real-world application. In practice, these two tasks are always mixed and combined to obtain a high-resolution enhanced (HRE) image with high quality from a low-resolution original image (LRO) with low quality. In this paper, we propose a joint SR-PIE learning framework called Deep SR-PIE, which comprises Multi-scale Backward Fusion Network (MBFNet), Perceptual Enhancement Network (PENet) and Dual-Path Unsampling Network (DUNet). MBFNet network is responsible for deep feature representation for further image reconstruction and perceptual enhancement, and PENet seeks the optimal local transformation to recover perceptual loss (color, tone, exposure and so on). DUNet works in different scales and exchanges each other to complement more details during upsampling. In our experiments, a real-world dataset is released to facilitate the development of joint learning for SR and PIE. Then, a thorough ablation study is provided to better understand the superiority of our method. Finally, extensive experiments suggest that the proposed method performs favorably against the state-of-the-arts in terms of visual quality, PSRN, SSIM, model size and inference time. By virtue of splitting operation and inverse residual blocks, as a lightweight deep neural network, our model is compatible with low-computation device.

INDEX TERMS Super resolution, perceptual image enhancement, lightweight.

I. INTRODUCTION

Image super-resolution and perceptual image enhancement are main research topics in the fields of computer vision and image processing. SR refers to recovering natural High Resolution (HR) images from Low Resolution (LR) images, which enjoys a wide range of real-world applications, such as medicine, public safety and surveillance, aerospace and so on. PIE typically learns the non-linear mapping from input images to retouched images, which is popular in photo retouching and computer vision [1]. There are some similarities and differences between these two tasks. For their similarities, they both attempt to reconstruct LR image with higher quality by alleviating blurs and artifacts. For their differences, SR cares more about detail reconstruction while PIE focuses more on perceptual enhancement (color rendition, light and contrast adjustment). Through a joint learning of SR and PIE, the visual quality of LR image can be enhanced in terms of contrast, color and detail with less side effects.

In our practical scenario, limited by unstable bandwidth and low computation capability, live streaming image are downscaled and retouched by our own Image Sensor Processor (IPS) and then displayed on terminal devices. In this way, we offer pleasing experience as well as decreasing network

The associate editor coordinating the review of this manuscript and approving it for publication was Seifedine Kadry.

delay. Afterwards, all the HR images are carried back with external disk or big hard disk, and then retouched by professional photographers with Adobe® Photoshop or Lightroom. Unfortunately, it takes arduous efforts to achieve pleasing visual results. Thus, it is essential to convert LR image to HR image with higher perceptual quality automatically.

In general, this joint problem is challenging and ill-posed since the pair of LR image and HR image is not a one-to-one correspondence. Recently, deep learning techniques have achieved substantial performance in various computer vision tasks and have greatly promoted the development of SR and PIE. To tackle SR task, a variety of deep learning methods based on traditional Convolution Neural Network (CNN) [2] and Generative Adversarial Nets (GAN) are developed. To handle PIE task, a serious of automatic methods are developed to address the issues of color rendition, image sharpness, brightness and contrast [1], [3]. For the joint problem, people takes for granted that producing HRE image from LRO one is to execute SR and PIE methods in sequence. Nevertheless, it is inefficient and inaccurate since error would be propagated in cascaded process. Moreover, when performing in joint scheme, the outputs generated by these two tasks could potentially complement each other to provide better results.

To address the joint SR-PIE problem, we consider it from a holistic perspective and then propose a deep neural network named Deep SR-PIE, which consists of Multi-scale Backward Fusion Network(MBFNet), Dual-Path Unsampling Network (DUNet) and Perceptual Enhancement Network (PENet). With respect to MBFNet, in order to increase the enlarge receptive field for extracting hierarchical features, we take different scales into account and adopt a backward concatenation for feature fusion. In detail, Multi-scale Splitting Block(MSB) is proposed to express multi-scale features, which retains partial information and proceeds other information with further layers. As for DUNet, it is proposed for upsampling through two-path shared convolution layers. With regard to PENet, an encoder-decoder network is proposed to learn the local transformation to correct the results from MBFNet. Benefiting from joint learning strategy, our method is capable to generate promising Super-Resolution Enhanced (SRE) images with low inference time. To better train our method, we release a real-world dataset named Alltuu2 involving 5K LRO and HRE image pairs.

Our contributions are summarized as follows: 1) We propose a novel multi-branch deep learning model for joint SR-PIE problem that achieves good trade-off between speed and performance. 2) We design an encoder-decoder network to learning the local transformation to improve the perceptual performance of high-resolution image. 3) We propose a new upsampling network with dual-path shared convolutions to enhance high-frequency details. 4) We contribute a new practical dataset named Alltuu2 for joint SR-PIE problem, and employ evaluation of the comparative methods on four datasets to figure out the superiority of our method.

The rest of this paper is organized as follows. Section II displays the discussion of related work. In Section III, the proposed method is described, and the experimental results are analyzed in Section IV. Finally, Section V concludes the paper and suggests possible topics for future research.

II. RELATED WORK

This work addresses the joint problem of SR and PIE for a single RGB image. In previous works, these two problems are well-studied separately. To our best knowledge, we are the first to solve the joint problem in a joint way. In this section, we first review several major works for Single Image SR (SISR) and PIE under supervision, respectively, and then discuss the literature of joint solution.

A. SUPER-RESOLUTION

Compared with traditional method, SISR methods based on deep learning show superior performance thanks to their powerful feature representation capability. With more recent structures employing residual block [4], dense connections [5] and channel attention blocks [6], most of the state-of-the-art approaches attempt to seek different optimal combinations of multiple modules. In residual learning, shortcut connections are used to directly connect input and output images(features). The pioneer work [2] proposes SRCNN to learn the mapping from LR image to HR image and achieve superior performance against previous works. Later, Kim *et al.* [7] and Zhang *et al.* [8] further improve SR performance by increasing the depth of CNN with residual learning. In SRResNet [9], stacked residual blocks and generative adversarial network are introduced to solve SR problem. Then, SRResNet is enhanced to a wider network EDSR and a deeper network MDSR via stacking more residual networks. In dense connections, for each layer in a dense block, the features maps of all preceding layers are treated as inputs, and their own feature maps are also passed to subsequent layers. Tong *et al.* [10] construct SRDenseNet with dense blocks and insert dense connection between different dense blocks to restore details and speed up SR model. To make full use of cleaner residual features, RFA [11] aggregates the local residual features with dense connection for more powerful feature representation. Whereas, the requirement of abundant parameters limits its application in real-time practice. MemNet [12], RDN [13], CARN [14] and DBPN [15] utilize wider and deeper network with layer-level and block-layer dense connections to boost their performance. Besides extensive efforts spent on designing wider and deeper structures, several models with attention modules are proposed to further enhance representation power of deep CNNs by exploring feature correlations along either spatial or channel dimension. RCAN [13] and SOCA [16] incorporate channel attention to further boost the performance. As an extension of channel attention, CS-NL [17] allows the network to concentrate on informative area with cross-scale feature correlation. However, its better SR results come at high computation cost.

B. PERCEPTUAL IMAGE ENHANCEMENT

Perception Image Enhancement has been studied for a long time in computer vision and image processing [3], [18]. Recently, with the success of CNN, pretty of learned-based approaches have been emerged as useful tools and enhance perceptual images in the following two aspects. The first one aims to learn a color-aware and lightness-aware mapping between the pairwise original-and-retouched training data. Yan *et al.* [19] automatically enhance color based machine-learned ranking approach. By means of local semantics in image, the automatic photo adjustment framework [20] applies stylistic color and tone adjustments to input original image after training on elaborated selected pairwise data samples. Gharbi *et al.* [21] introduce bilateral learning by combining bilateral grid algorithm with local affine color transformation for real-time application. Besides, reinforcement learning is also incorporated to generate understandable operation sequences for photo retouching [22], [23]. The second one mainly pays regard to detail recovery of low-quality image. Hybrid model is designed to model high-frequency edges and low-frequency image contents using Recurrent Neural Network (RNN) and autoencoder technique, respectively [24]. To compensate for lost details, Enhancenet [25] generates images with more realistic textures by using a perceptual loss, and reference [26] learns a translation mapping from ordinary photos to DSLR-quality photos to improve both rendition and image sharpness. Moreover, the issue of low-light image enhancement is similar to perceptual image enchantment. Apart from the factors of low-light and noise, perceptual image enhancement also focuses on the color variant. Lore *et al.* [27] make the first attempt by training Low-Light Net (LLNet) on random Gamma correction for contrast enhancement and noise removal. Later on, various methods on more complex networks are proposed with paired dataset [24], [28] or unpaired dataset [29], [30]. Because these method are oriented towards enchaining contrast information rather than color change, the results of these methods are still imperfect especially failing to restore color variance.

C. JOINT SR-PIE

In real-world application, it is impossible to remain image perceptual quality when resolving SR problem. In the field of joint SR-PIE, much effort is paid for raw images rather than RGB images. For 12-bit or 14-bit raw images, a series of approaches analyze intrinsic mechanism of raw image and learn a large collection of operations (e.g., demosaicing, denoising, compression and color correction) to approximate nonlinear ISP pipelines. To be concrete, Schwartz *et al.* [31] resort to deep CNNs for learning color correction mapping of specific digital cameras. Following their work, Xu *et al.* [32] design a dual network to exploit both raw data and color image for real scene super-resolution, which generalizes well to different cameras. Also, HERN [33] employs two parallel paths to learn image features in two different resolutions. For joint SR-PIE problem, recent methods only treat PIE as auxiliary product when solving SR problem for raw images, and most of them only concern about details rather than color. However, the mixture problem of SR and PIE has not witnessed jointly learning strategy to the best of our knowledge.

III. METHODOLOGY

Many recent SR networks have similar network structure, similar to several state-of-the-art methods [11], [13], the overall pipeline of our proposed framework is depicted in Fig. 1, which is mainly divided into three components: Multi-scale Backward Fusion Network (MBFNet), Dual-path Unsampling Network (DUNet) and Perceptual Enhancement Network (PENet). In order to reduce computation and spatial complexity, we prefer putting upsampling operation at the end. The detail of the proposed framework is elaborated below.

$$I^{MBFNet} = H_{MBFNet}(I, I_{i+d}) \tag{1}$$

$$I^{PENet} = H_{PENet}(I^{MBFNet}, I) \tag{2}$$

$$I^{SRE} = H_{DUNet}(I^{PENet}) \tag{3}$$

where I^{MBFNet} and I^{PENet} are the outputs of MBFNet and PENet. $H_{MBFNet}(\cdot)$, $H_{PENet}(\cdot)$ and $H_{DUNet}(\cdot)$ denote the network of MBFNet, PENet and DUNet, respectively. MBFNet network is responsible for deep feature representation for further image reconstruction and perceptual enhancement. PENet network is devised to estimate local transformation on above deep features I^{MBFNet} and input image I. To achieve SR objective, DUNet network is designed for image reconstruction where dual-path shared convolutions followed by pixelshuffle modules are applied. I_{i+d} means the concatenation of I and I_d. I^{SRE} is the final output of Deep SR-PIE.

A. INPUT DECOMPOSITION

Inspired by the work [34], an efficient and effective guided filter is used to preserve edges and textures. For a given LRO image I and its corresponding HRE image I^{HRE}. In our pipeline, I is first decomposed into the base layer I_b and the detail layer I_d. I_b is obtained by applying low-pass filter to LRO image, and I_d is calculated by simply diving I by I_b.

$$I_b = low\_filter(I) \tag{4}$$

$$I_d = I \oslash I_b \tag{5}$$

where $\oslash$ is element-wise division operation and *low_filter* denotes low-pass filter. I_d is dominant with high frequency information that can preserve edges and textures. Owning to the specific usages of the detail layer, we entitle our model to describe high frequency more powerfully by using I_{i+d} as one of its input, which is formulated as:

$$I_{i+d} = I + I_d \tag{6}$$

B. MULTI-SCALE BACKWARD FUSION BLOCK

As illustrated in Fig. 1, MBFNet mainly consists of three paths: Multi-scale Splitting Block (MSB), Backward

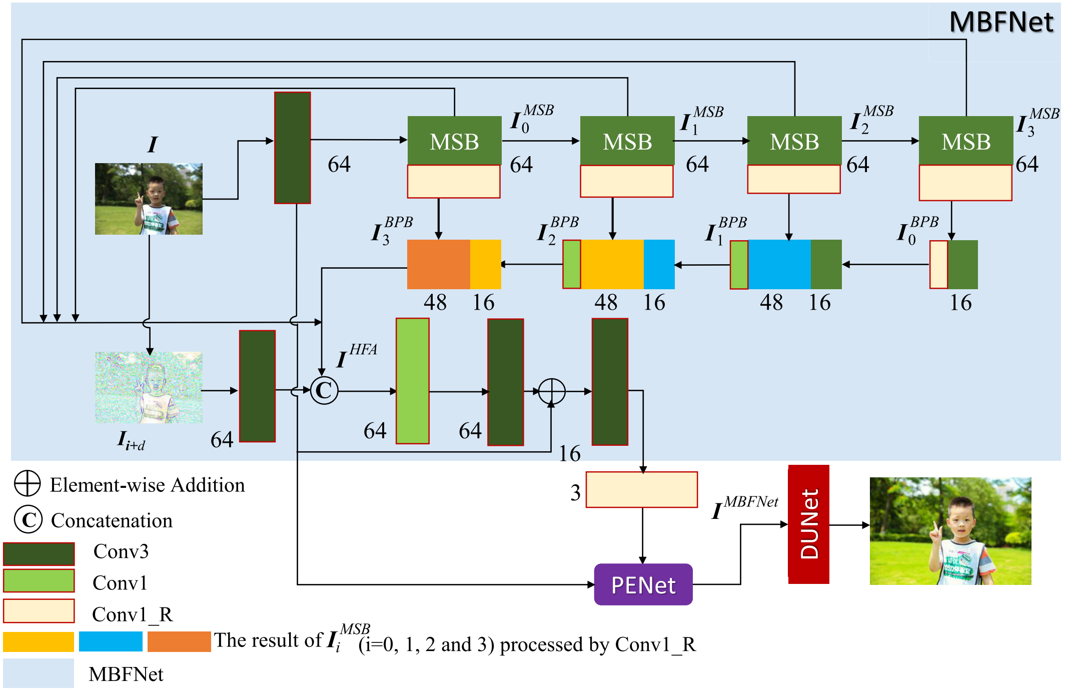


FIGURE 1. The overall pipeline of our proposed framework Deep SR-PIE.

Propagation Block (BPB) and Hybrid Feature Aggregation (HFA).

$$\begin{aligned} I_t^{MSB} &= H_{MSB}(I_{t-1}^{MSB}) \\ &= H_{MSB}(H_{MSB}(\cdots(H_{MSB}((\mathcal{H}(I))\cdots)) \quad (7) \\ I_{k+1}^{BPB} &= H_{BPB}(I_{t-k-1}^{MSB}, I_k^{BPB}) \quad (8) \\ I^{HFA} &= Concat(I_0^{MSB}, I_1^{MSB}, \cdots, I_t^{MSB}, I_t^{BPB}, \mathcal{H}(I_{i+d})) \\ & \quad (9) \\ I^{MBFNet} &= H_{MBFNet}(I, I_{i+d}) \\ &= Conv3(Conv3(Conv1\_R(I^{HFA})) \oplus \mathcal{H}(I)) \quad (10) \end{aligned}$$

where H_{MSB}, H_{BPB} represents the modules of MSB and BPB, respectively. I_k^{MSB} and I_k^{BPB} respectively denote the outputs of the k-th MSB and BPB($k \in [0, t]$). Given the input, we can get the shallow feature $\mathcal{H}(\cdot)$ where $\mathcal{H}$ stands for the convolution operator 3×3 in our implementation. $Conv1\_R(\cdot)$ is convolution 1×1 followed by RRelu activation and $Conv3$ is 3×3 convolution.

1) MULTI-SCALE SPLITTING BLOCK

The goal of MSBs is to extract features for deep feature learning. Inspired by Inception block [35] and channel splitting idea [36], we design Multi-scale Splitting Block (MSB) and utilize $t + 1$ MSBs to capture the features at different scales. As illustrated in Fig. 2, MSB uses a series of multi-scale residual splitting operations to extract different scales features. Firstly, as dilated convolution [37] could increase the receptive field under the condition that the resolution of feature map is unchanged, we use a dilated convolution 3×3 followed by RReLu layer to perceive more information. Then, we put forward a series of splitting steps to produce multi-scale features efficiently. For the first step, MSB employs 1×1, 3×3 and 5×5 convolution layers to split the preceding features into three parts. The result of 1×1 convolution layer is retained, and the other two parts are fed into the next steps. For the second step, the result of 3×3 convolution layer is split into two parts with channel splitting operation. One is preserved and the other part is fed into next step. Also, the above procedure is applied to the result of 5×5 convolution layer. After two splitting steps, all the distilled features are concatenated together and then fed into a 1×1 convolution to reduce the channels and parameters.

In SR task, the reconstruction performance depends on the perception of high-frequency features. Accordingly, Frequency-aware Channel Attention (FCA) is proposed to select frequency-aware contextual information. Let x_{con} is the above concatenated features, we first adopt two dilated convolutions to extract features in different receptive fields. The two convolutions f_{d1} and f_{d2} are $kernel = 1$ with dilation $rate = 1$ and $kernel = 3$ with dilation $rate = 2$, respectively. Next, we consider a high-frequency ratio map C_r between these two groups.

$$C_r = sigmoid(f_{d1}(x_{con}) - f_{d2}(x_{con})) \quad (11)$$

where C_r is the pixel-wise contrast information where the pixel with high contrasts, and $sigmoid$ is sigmoid function. Then, we multiply x_{con} by C_r to obtain high-frequency map $C_h = C_r \cdot x_{con}$. Finally, the output of FCA layer is expressed as $I^{FCA} = C_h \oplus C_r$ that acts as the enhancement of x_{con}. With the assistance of FCA module, our network can improve the SR

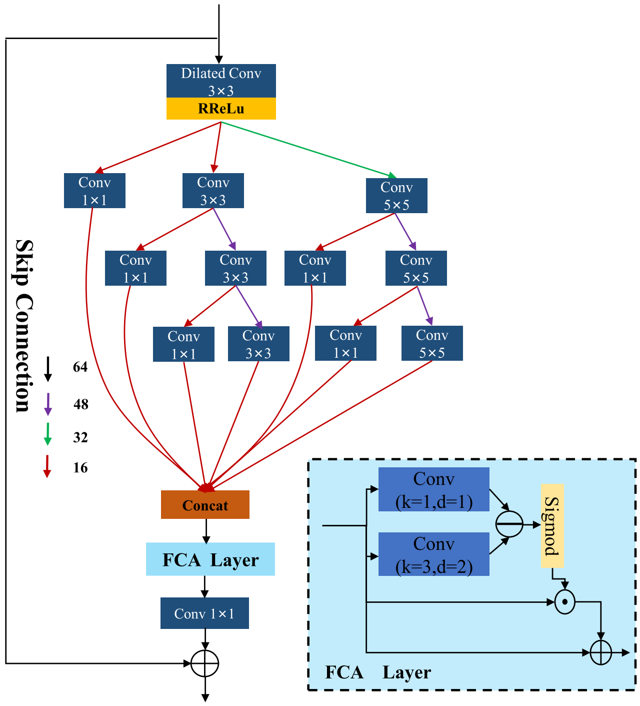


FIGURE 2. The architecture of MSB. Here, 64, 48, 32 and 16 represent the output channels of the convolution layers. 'FCA layer' indicates the proposed frequency-aware channel attention (FCA).

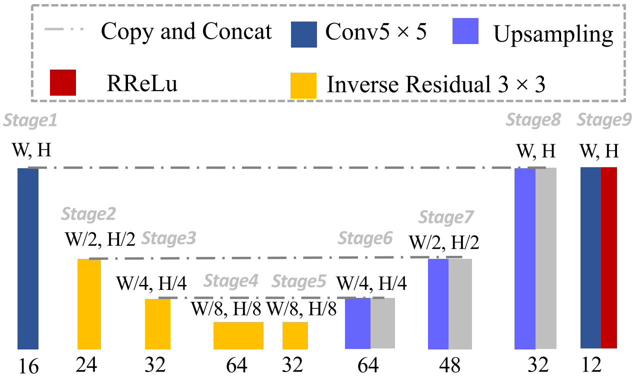


FIGURE 3. The details of perceptual enhancement network.

performance steadily. Besides, a skip connection is utilized to enhance the feature propagation of I^{FCA}, yielding the output of $I_{k+1}^{MSB} = I^{FCA} \oplus I_k^{MSB}$.

2) BACKWARD PROPAGATION BLOCK

As we all know, hierarchical information plays an importance role in reconstructing image. Therefore, we present $t + 1$ backward propagation blocks to fuse the features from the preceding MSBs. As illustrated in Fig. 1, there are $t = 3$ BPBs. Mathematically, the output of the $(k+1)$-th BPB I_{k+1}^{BPB} is formulated as:

$$I_{k+1}^{BPB} = \begin{cases} Conv1\_R(I_{t-k-1}^{MSB}) & if\ k == 0, \\ Concat(Conv1(I_{t-k-1}^{MSB}), Conv1(I_k^{BPB})), & otherwise \end{cases} \tag{12}$$

where *Concat* is the concatenation operation. From Eq. 12, except the first BPB, the feature I_k^{BPB} followed by a 1×1 convolution is concatenated with I_{t-k-1}^{MSB} followed by a 1×1 convolution. In this way, we find such a backward fused strategy helps to integrate the features with more hierarchical contextual information.

3) HYBRID FEATURE AGGREGATION

As a aggregation module of all the features from different branches, HFA is able to provide a more representative feature for SR task. As displayed in Eq. 9, HFA is designed to concatenate the following hybrid features. Regarding the first branch HFA_1, the output of I_t^{BPB} is generated subsequently to make a better use of hierarchical features. As the second branch HFA_2, the output of the t-th MSB is sent directly to the end of I_{k+1}^{BPB}, which link the MSBs with I_{k+1}^{BPB} in order to provide rich information for SR reconstruction. For the last branch HFA_3, the input I_{i+d} carried with high-frequency details pass through a 3×3 convolution to compensate the lost high-frequency information. After that, a concatenation operation is applied to stack the above features together termed as I^{HFA}. Compared with the way of simply stacking multiple residual skipping, HFA attempts to ensure that the useful information can be propagated to the next layer without any significant loss, leading to a more discriminative feature representation.

To alleviate the vanishing-gradient problem, a skip connection from the shallow feature of input I is introduced into MBFNet. As illustrated in Eq. 10, $Conv3(Conv1\_R(I^{HFA}))$ is element-wise addition with and $\mathcal{H}(I)$, and then pass through 3×3 convolution to yield the output of MBFNet I^{MBFNet}.

C. PERCEPTUAL ENHANCEMENT NETWORK

As another critical branch, PENet and its postprocessing are designed to correct and restore lost perception between LRO and HRE. PENet aims to learn a pixel-wise mapping parameters for local transformation, and then is applied to the preceding mapping on I^{MBFNet} in order to enable more versatile adjustment. The input of PENet is the shallow feature $\mathcal{H}(I)$ while the output is local transformation. Detailedly, PENet has an encoder-decoder structure. The encoder reduces the spatial resolution of $\mathcal{H}(I)$ to exploit larger receptive fields for spatial filtering, while the decoder performs upsampling to restore the spatial resolution. The detail of PENet is depicted in Fig. 3. Given the input shallow feature $\mathcal{H}(I)$, it is encoded and gradually downsampled with a series of inverse residual 3×3 convolution operations until $\frac{W}{8} \times \frac{H}{8} \times 32$. In decoder part, upsampling denotes the deconvolution operation to increase the size of the feature map with scale ×2. 'Copy and Concat' copies the outputs of 'Stage 1', 'Stage 2' and 'Stage 3', and concatenate with the previous results in 'Stage 6', 'Stage 7' and 'Stage 8'. After three upsampling stages with 'Copy and

Concat', the feature map shares identical size with the input is obtained. In 'Stage 9', we utilize 1 × 1 convolution to reduce the channel dimension and output the $\mathcal{A}_{PENet}$ with the size of $W \times H \times 12$.

Motivated by the works [31], [38], [39], we try to learn a pixel-wise local transformation to recover perceptual loss. Since global enhancement adjustment methods work for all the pixels, they always over-/under-enhance local regions in most cases. To address this problem, as a postprocessing step, we resort to matrix $\mathcal{A}_{PENet}$ with the size of $W \times H \times 12$ as a pixel-wise local transformation for each pixel, and write it as follows:

$$\begin{aligned} I^{SRE} = &\sum_{i=0}^{2} \mathcal{A}_{PENet}[:, :, i \times 3 : (i+1) \times 3] \\ &\times I^{MBFNet}(1 - I^{MBFNet}) + \mathcal{A}_{PENet}[:, :, 9 : 11] \\ &\times I^{MBFNet} \end{aligned}$$

D. DUAL-PATH UPSAMPLING NETWORK

As an important end-to-end learnable layer, pixelshuffle has been widely used to upsample in SISR task [40]. Different from deconvolution layer, it generates a series of channels by convolution and then reshape them. After this layer, a feature map can be upsampled with s^2 times channels where s is the scaling factor. However, limited by fixed kernel size, pixelshuffle fails to provide enough contextual information to restore realistic details. In order to address the problem, we propose Dual-path Upsampling Network (DUNet) which works in different scales and exchange each other to complement more details. As illustrated in Fig. 4 and Eq. 13, the input $\mathbf{F}^{MSDAB}$ with 3 channels is processed by 3 × 3 and 5 × 5 convolutions, respectively. Using the channel splitting strategy in Section III-B1, the preceding results are split into three parts. the first (FT_{1_1} or FB_{1_1}) goes forward the next convolution layer, the second (FT_{1_2} or FB_{1_2}) moves to another path, and the last (FT_{1_3} or FB_{1_3}) acts as skip connection. During these two-bypass convolution operations, the information between those bypasses can be shared each other. Then, they are passed through shuffle operations to upsample s^2 times. Finally, all of these feature maps are concatenated and sent to a 1 × 1 convolution layer. Formally, the above procedure can be written as:

$$\begin{aligned} FT_1 &= Conv5(\mathbf{F}^{MSDAB}) \\ FB_1 &= Conv3(\mathbf{F}^{MSDAB}) \\ FT_{1\_1}, FT_{1\_2}, FT_{1\_3} &= Split(FT_1) \\ FB_{1\_1}, FB_{1\_2}, FB_{1\_3} &= Split(FB_1) \\ FT_2 &= Conv5(Concat(FT_{1\_1}, FB_{1\_2})) \\ FB_2 &= Conv3(Concat(FB_{1\_1}, FT_{1\_2})) \\ FT_3 &= Shuffle(Concat(FT_2, FT_{1\_3}))) \\ FB_3 &= Shuffle(Concat(FB_2, FB_{1\_3})) \\ I^{SRE} &= Conv1(Concat(FT_3, FB_3)) \end{aligned} \tag{13}$$

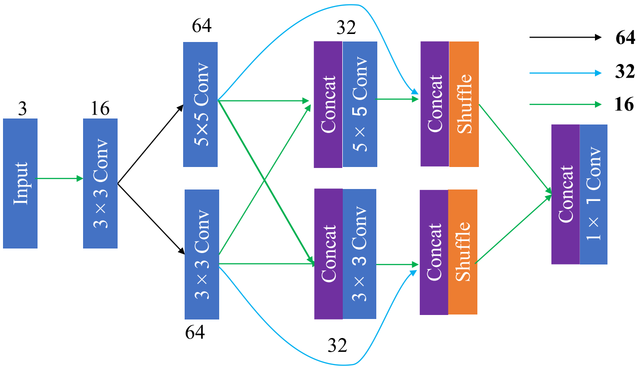


FIGURE 4. The architecture of dual-path upsampling network.

where *Split* and *Shuffle* denote channel splitting operation and PixelShuffle operation, respectively.

E. HYBRID LOSS AND EVALUATION METRICS

1) LOSS FUNCTION

Apart from network architecture, loss function also plays a key part in network design. In this paper, we propose a hybrid loss function containing five components and minimize it during our training. Given an input image I, the predicted SRE image I^{SRE} and the ground-truth image I^{GT}, the hybrid loss can be expressed as Eq.14.

$$\mathcal{L} = w_1\mathcal{L}_{con} + w_2\mathcal{L}_{tv} + w_3\mathcal{L}_{color} + w_4\mathcal{L}_{MSSIM} + w_5\mathcal{L}_1 \tag{14}$$

where $\mathcal{L}_{con}$, $\mathcal{L}_{tv}$, $\mathcal{L}_{color}$, $\mathcal{L}_{MSSIM}$ and $\mathcal{L}_1$ are loss components, and w_1, w_2, w_3, w_4 and w_5 are their corresponding tunable weights, respectively. Empirically, we set $w_1 = 0.001$, $w_2 = 1$, $w_3 = 0.0005$, $w_4 = 300$ and $w_5 = 0.05$.

Content Loss: Content loss, also named as perceptual loss, makes many contributions on SR and PIE. The goal of content loss is to encourage distances of images in feature representations to be as close as possible. In our case, it helps to preserve image semantics to some extent. Generally, the feature space constructed by an pre-trained VGG-19 model proves effective in previous works [41]. So, content loss is defined as:

$$\mathcal{L}_{con} = \frac{1}{C_j H_j W_j} \left\| \varphi_j(I^{SRE}, I^{GT}) \right\| \tag{15}$$

where C_j, H_j, and W_j are the number of channels, height, width of feature maps obtained by j-th convolutional layer of VGG-19 CNN network. Here, we use pool4 layer as feather extractor when training our framework.

Total Variation Loss: Typically, it is observed that networks only with content loss are inclined to generate highly pixelated and noisy output. Consequently, we add total variation loss to enforce spatial smoothness and continuity. Intrinsically, total variation loss can be treated as regularization loss.

$$\mathcal{L}_{tv} = \frac{1}{CHW} \left\| \nabla_x I^{SRE} + \nabla_y I^{GT} \right\| \tag{16}$$

where C, H, W are the size of output I^{SRE}.

Color Loss: In addition to content loss and total variation loss, differences in contrast, brightness and color are considered to encourage the above properties of I^{SRE} to match those in I^{GT}. For the purpose of enhancing perceptual quality, color loss is incorporated to solve our joint SR-PIE problem.

$$\mathcal{L}_{color} = \left\| I_{dk}^{SRE} - I_{dk}^{GT} \right\|_2^2 \tag{17}$$

where I_{dk}^{SRE} and I_{dk}^{GT} are the blurred images of I^{SRT} and I^{GT}, respectively. In contrast to Gaussian blur, dual kawase blur is more efficient to produce comparable or even better results.

MSSIM Loss: As one of our goals, our framework attempts to produce visual pleasing images. The networks with structural similarity index (SSIM) loss are always encountered with the dilemma of whether set larger Gaussian coefficient or smaller one [42]. Rather than fine-tuning the coefficients, a Multi-scale SSIM (MSSIM) is utilized to extract SSIM at different scales based on the sensitivity of HVS. Given a pixel p from the same spatial location from I^{SRE} and I^{GT}, its MSSIM can be written as

$$MSSIM(p) = [l_M(p)]^{\alpha_M} \cdot \prod_{j=1}^{M} [c_j(p)]^{\beta_j} [s_j(p)]^{\gamma_j} \tag{18}$$

where $l_M(p)$ is the luminance comparison at scale M. At j-th scale, $c_j(p)$ and $s_j(p)$ denote the contrast comparison and the structure comparison, respectively. α_M, β_j and γ_j are used as the weights of different components. For convenience, in most cases, $\alpha_M = \beta_j = \gamma_j = 1$. To alleviate the difficulty of calculating the derivatives, an alternative MSSIM can be found.

$$\mathcal{L}_{MSSIM} = 1 - MSSIM(\tilde{p}) \tag{19}$$

where $\tilde{p}$ is the center pixel of input images. As proved in [42], the network kernels learned by the center pixel can be also equally applied to the other pixel in the images.

$\mathcal{L}_1$ *Loss* Pixel-wise loss is critical to measure reconstruction error and guide model optimization. Instead of $\mathcal{L}_2$ loss, the networks equipped with $\mathcal{L}_1$ loss tend to learn better signal-to-noise ratio (PSNR) [43], [44] that is highly correlated with pixel-wise difference. Thus, we also prefer $\mathcal{L}_1$ loss to avoid getting stuck in a local minimum.

$$\mathcal{L}_1 = \left\| I^{SRE} - I^{GT} \right\|_1 \tag{20}$$

2) EVALUATION METRICS

Evaluation metrics play an import role in measuring the performance of models. As we all know, there is no unified and admitted metric to evaluate image quality objectively. To be fair, we also adopt the most widely used metrics PSNR and SSIM as our evaluation metrics. With regard to PSNR, it is used to measure signal distortion between I^{SRE} and I^{GT}, whereas it would result in incredible value even if two image are almost indistinguishable. In reference to SSIM, it focuses on measuring the perceptual quality of brightness, contrast and structure. As for these two common metrics, the higher are the values, the better are the quality of I^{SRE}.

IV. EXPERIMENTS

In this section, we conduct our experiments to evaluate the performance of our proposed framework for joint SR-PIE problem. The experiments include three parts: The first part makes ablation studies of our framework, the second part evaluates our proposed framework against several state-of-the-art methods on various benchmark datasets and the last part presents some failure cases.

A. DATASETS

1) TRAINING SETS

Following the previous works, we choose DIV2K dataset as one of our training sets. It contains 800 LR-HR 2K resolution images and spans a variety of image categories, including animal, building, food, landscape, people, plant, etc. We use all the 800 high-resolution images as training images. It was collected for NTIRE2017 and NTIRE2018 Super-Resolution Challenges in order to encourage research on image super-resolution with more realistic degradation. As illustrated in Fig. 5(a,b), no perceptual loss is found in dataset DIV2K. Note that there is no specific benchmark dataset for the joint problem of SR and PIE, so we further release a real-world dataset called Alltuu2 which is captured with various ISP equipments (Canon EOS 5D Mark IV, NIKON D810, Canon EOS 5D Mark III, etc.). In our practical application, millions of 2K+ images from different scenes are captured and stored according to scene category. For these LRO images, several skilled photographers enhance the images with Adobe® Photoshop or lightroom, yielding LRO-HRE image pairs. Then, we randomly extract image pairs from different categories to avoid high coherence among these images. As seen in Fig. 5(c,d), resolution loss and perceptual loss are found in dataset Alltuu2. In total, 5,153 training images are randomly selected from Alltuu2 in our experiments. All relevant codes and our own dataset are available in Deep SR-PIE soon.

2) TESTING SETS

After treating DIV2K dataset as training set, we adopted four widely-used super resolution benchmark datasets in inference procedure. BSD100, as a subset of BSD 500 [45], provides 100 natural scenes collected from Berkeley segmentation dataset. Set14 and Set5 consists of 14 images and 5 images with different objects reported in [46], respectively. Urban100 contains 100 HR images with a variety of real-world building structures that fetched from Flickr using keywords such as urban, city, architecture, and structure [47]. In the light of these four datasets, their heights or widths range from 228 to 566. As a matter of convenience, the above four benchmark datasets are collectively named dataset SRT. For Alltuu2, 304 images are prepared for testing.

B. IMPLEMENTATION DETAILS

1) TRAINING DETAILS

To boost the performance and generality ability of our model, random horizontal and vertical flipping(random probability

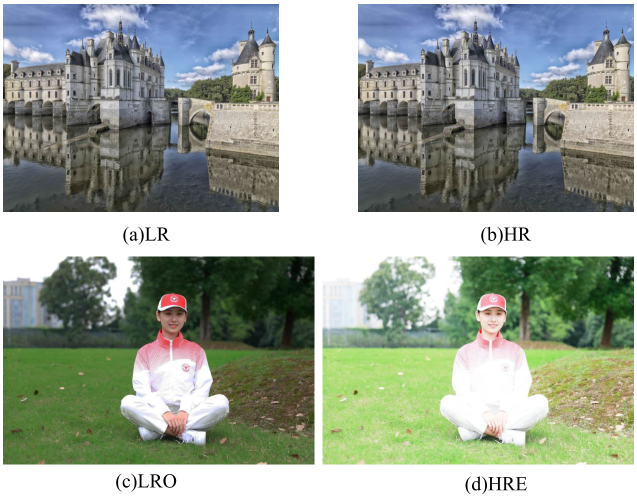


FIGURE 5. **The examples of training images. (a) (b) An pair example (LR and HR) in dataset DIV2K. (c) (d) An pair example (LRO and HRE) in dataset Alltuu2.**

is 1) are used for data argumentation. In total, there are 1,600 and 10,306 training images in DIV2K and Alltuu2, respectively. The proposed network is implemented with Pytorch 1.1.0. The network is trained on a single NVIDIA® 1080Ti GPU by optimizing loss $\mathcal{L}$ for 120 epochs with varied batch sizes. Inspired by the progressive training strategy used in [33], we grow the resolutions of input and ground-truth gradually. For the first 50 epoch, we use 64 × 64 patches and set the learning rate $lr = 1 \times 10^{-4}$. For the last 70 epoch, we use 88 × 88 patches and set the learning rate $lr = 1 \times 10^{-5}$. With different patch sizes during training, the batch size is also decreasing from 16 to 4. Adam optimizer with setting ($\beta_1 = 0.9$, $\beta_2 = 0.999$) is adopted. To prevent overfitting, we set dropout rate to 0.5. To be fair, all the comparative methods are implemented and line with the hyperparameters and parameters in their papers.

2) TESTING DETAILS

In this section, self-ensemble strategy is applied on testing images to improve model performance and robustness. Concretely, three different operations, horizontal flipping, vertical flipping and horizontal-vertical flipping are carried out on testing image. Four different images including original one are fed into our framework to get a set of four temporal images. Then, the corresponding inverse transformations are executed on the temporal set to produce the outputs. For the final prediction image, it is conducted by the average of these outputs.

In addition, we also evaluate the generalization ability of our model. Fig. 6 shows the training loss curve vs testing loss over the number of the epochs on dataset Alltuu2. As can be seen, as the number of the epoch increases, the gap between training and testing loss is stable. Namely, our model has strong generalization ability.

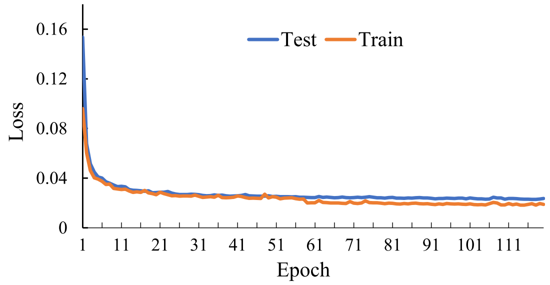


FIGURE 6. **The training loss curve vs testing loss over the number of the epochs on dataset Alltuu2.**

C. ABLATION STUDY

To better evaluate our proposed framework, we conduct ablation studies by removing each component and keeping the rest unchanged. We place particular emphasis on differences brought by four main components. Thus, we design and analyze the following experiments.

1) STUDY OF INPUT COMPOSITION

In this section, we validate the effectiveness of different combinations of the inputs. In MSB, it requires more complete and precise image information while either I_b or I_d will bring color cast. Thus, we only treat I as the input of MSB and employ no further experiment on it. In the last branch of HFA(HFA_3), we take three different combinations of input composition into account, and place their results in Table 1(1-4). Notable benefit is gained with stacking I and I_{i+d} that conforms with the report in [34]. Compared with the whole image I, I_{i+d} puts particular emphasis on high-frequency and detailed information. In addition, no further performance is gained by I_b since the features only dominated with low-frequency information are deficient to reveal enough discriminated ability. Aside from quantitative analysis, an visual qualitative comparison of HFA_3 with the above settings is placed in Fig. 7. From the perspective of information utilization, we can inform that I_{i+d} better preserves more informative high-frequency details and depress noise and artifact at the same time, which is in accord with the goal of HFA_3.

2) STUDY OF MBFNet

In this section, the effectiveness of critical operations will be investigated. At first, we make a thorough comparative experiment to verify the importance of BPB. In general, BPB fuses feature from multiple layers to obtain more contextual information. As displayed in Table 1(1, 5), after removing HFA_1 from our model, it is found the performance drops 5.87% and 1.8% on metric PSRN and SSIM, respectively. From the observation, it indicates BPB has a significant influence on the performance. Then, we continue to verify the superiority of skip connection from the shallow feature of input I. Skip connection is always suggested in deep network architecture to preserve long-term memory for residual learning. When

TABLE 1. Ablation results (PSNR / SSIM) of the Deep SR-PIE model on dataset Alltuu2. The best results are highlighted in bold.

No	Input				MBFNet			PENet			DUNet	PSNR	SSIM
	$(\mathbf{I}_{i+d}, \mathbf{I})$	$(\mathbf{I}_d, \mathbf{I}_b)$	$(\mathbf{I}, \mathbf{I})$	$(\mathbf{I}_d, \mathbf{I})$	HFA_1	HFA_2	HFA_3	$\mathcal{A}_{PENet}$	$\mathcal{A}_{v1}$	$\mathcal{A}_{v2}$	PixelShuffle		
1	✓				✓	✓	✓	✓				**29.8379**	**0.9649**
2		✓			✓	✓	✓	✓				29.6535	0.9609
3				✓	✓	✓	✓	✓				29.6946	0.9617
4			✓		✓	✓	✓	✓				29.7498	0.9622
5	✓					✓	✓	✓				28.1826	0.9478
6	✓				✓		✓	✓				29.2153	0.9604
7	✓				✓	✓		✓				28.8468	0.9557
8	✓				✓	✓	✓		✓			29.5842	0.9609
9	✓				✓	✓	✓			✓		27.7855	0.9467
10	✓				✓	✓	✓					28.7678	0.9613
11	✓				✓	✓	✓	✓			✓	29.4711	0.9598

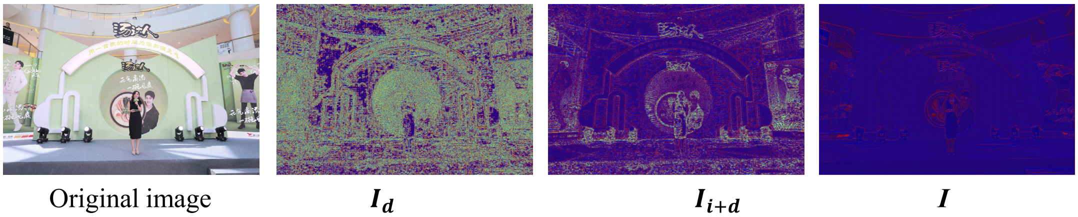
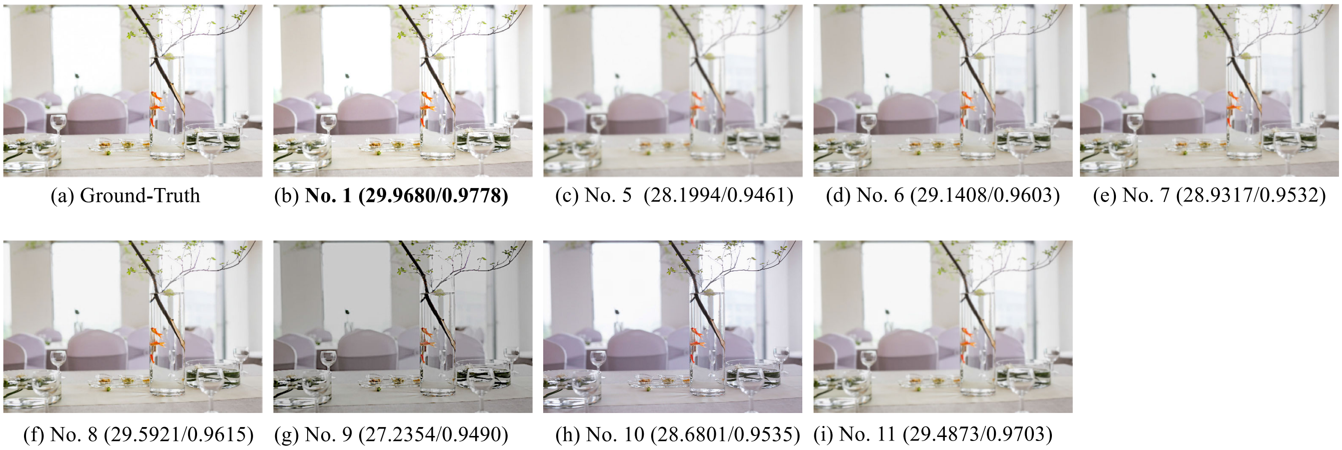


FIGURE 7. The visual qualitative result of different input of the last branch of HFA(HFA_3).


(a) Ground-Truth (b) **No. 1 (29.9680/0.9778)** (c) No. 5 (28.1994/0.9461) (d) No. 6 (29.1408/0.9603) (e) No. 7 (28.9317/0.9532)

(f) No. 8 (29.5921/0.9615) (g) No. 9 (27.2354/0.9490) (h) No. 10 (28.6801/0.9535) (i) No. 11 (29.4873/0.9703)

FIGURE 8. The visual ablation comparative results with metrics PSNR and SSIM.

removing the skip connection(HFA_2), Table 1(6) shows the quantitative ablation results, from which we can find slight performance decreases comparing with our model. Finally, we evaluate the model when (HFA_3) is eliminated. In this case, no high-frequency and detailed information is used to compensate the proposed model. As seen in Table 1(1, 7), HFA_3 plays an active role in the reconstruction process of SR task. Moreover, we place their visual results in Fig. 8. From Fig. 8(c), with the same settings in perceptual enhancement, the difference between them and the ground-truth is the quality of reconstructed details. Without BPB models, the model produces unnatural and blurry texture. From Fig. 8(b,d,e), we find that the removal of HFA_2 or HFA_3 will result in unclear details. As a summary, the above analysis demonstrate that HFA_1, HFA_2 and HFA_3 are of crucial importance for fully exploiting deep features and high-frequency detail from LRO images.

3) STUDY OF PENet

To investigate the importance of PENet, we design a series of experiments with different structures. As mentioned in

Section III-D, we apply another two local transformations to correct the perceptual loss of I^{SR}. For the first implementation, we refer to [48] and name it as $\mathcal{A}_{v1}$. Its definition can be formulated as follows:

$$I^{SRE} = \mathcal{A}_{v1} \times I^{MBFNet}(1 - I^{MBFNet}) + I^{MBFNet} \tag{21}$$

Similar to $\mathcal{A}_{PENet}$, $\mathcal{A}_{v1}$ is learned by PENet in which the convolution of 'Stage 9' is substituted with 5×5 convolution with 3 channels. For the second implementation, we consider the process in [39] and revise it as $\mathcal{A}_{v2}$.

$$\begin{aligned} I^{SRE} &= \mathcal{A}_{v2} \times I^{MBFNet}(1 - I^{MBFNet}) + I^{MBFNet} \\ \mathcal{A}_{v2} &= S(x) \end{aligned} \tag{22}$$

where x means the position of a pixel, and $S(x) \in [-1, 1]$. For the pixels with different light conditions (overexposure or underexposure), the functions $S(x)$ are totally different. To achieve a clear result with rich details, we take brightness and gradient information into account. We first choose a moderate bezier curve whose coordinates are [0,1,1,1] to get a illumination adjustment result V_h. V_b is taken as the borderline of I^{MBFNet}. Then, we calculates the gradient of V_h and V_b, and get their difference $D(x) = \nabla V_h - V_b$. Next, we employ OTSU on D to get a threshold t_g.

$$S(x) = \begin{cases} D(x), & D(x) \geq t_g, \\ 0, & D(x) \leq t_g \end{cases} \tag{23}$$

The quantitative comparison results are given in Table 1 (1, 8, 9, 10). Local transformations ($\mathcal{A}_{PENet}$, $\mathcal{A}_{v1}$) share the same processing pipeline which applies the output of *PENet* to perceptual enhancement. In this way, I^{SR} are mapped to the ground-truth better. The biggest difference between these two local transformations is the postprocessing procedure. Instead of the linear combination of several matrices in postprocessing, the transformation in $\mathcal{A}_{v1}$ for each channel is a given matrix. Also, the coefficient of addition term I^{MBFNet} is set to 1. Note that $\mathcal{A}_{v2}$ is independent on the output of PENet, which is determined by the brightness and gradient information of I^{MBFNet}. As shown in Table 1 (1, 8, 9, 10), our method performs better than the others. In Fig. 8, we observe that the sample processed by $\mathcal{A}_{v2}$ (Fig. 8 (g)) is color cast, and the model without $\mathcal{A}_{v1}$ (Fig. 8 (f)) is the most visually similar to the ground-truth. In order to show the necessity of branch PENet, we train our model where branch PENet has been removed. From Fig. 8 (h) and Table 1 (10), it can be seen that high-fidelity color cannot be satisfactorily recovered in this way.

4) STUDY OF DUNet

We conduct ablation analysis to concern about the contribution of *DUNet*. As a contrast, the comparative module is constructed with pixelshuffle and trained on our own dataset. In table 1 (1, 11) and Fig. 8 (i), DUNet helps our model to obtain better PSRN and SSIM by ways of shared features brought by two-bypass convolution operations described in Section III-D.

D. PERFORMANCE EVALUATION

1) COMPARATIVE METHODS

To our best knowledge, there is no public research for joint SR-PIE problem. So far, there are enormous state-of-the-art works toward SR and PIE independently. In some excellent researches, latent key points in downscaled image are perceived and explored to boost SR performance in the subsequent upsampling procedure [49]. Nevertheless, the donwsampling operation is always unknown or even nonexistent in real-world application. In our experiments, all of those methods are not taken into consideration. Yet for all that, we still try to train several SR methods and give them the ability to address joint SR-PIE problem.

CS-NL: The cross-scale nonLocal attention module (CS-NL) [17] is proposed to sufficiency discover the widely existing cross-scale feature similarities in nature images. It is then integrated with local and previous in-scale non-local priors to benefit SISR.

EDRN: The encoder-decoder residual network (EDRN) [50] introduces an encoder-decoder structure with coarse-to-fine scheme. By means of larger receptive field and meticulous network design, it can describe features with more convex information and restore lost information gradually. Also, it also makes a thorough discussion whether the usage of batch normalization is efficient or not in real-world SISR problem.

IMDN: The lightweight information multi-distillation network (IMDN) [36] solves SISR problem by constructing distillation and selective fusion parts. It conducts a progressive refinement module to extract hierarchical feature and steadily improves SR performance with contrast-aware channel attention. To be compatible with real-world images with arbitrary size, it presents a adaptive cropping strategy.

RFA: The residual feature aggregation (RFA) groups the residual block together along with useful hierarchical features [11]. Besides, an enhanced spatial attention (ESA) is introduced to focus on spatial contents of key importance. The final network is constructed by applying RFA with the ESA blocks, which produces comparable SR results.

RCAN: The very deep residual channel attention Networks (RCAN) [51] proposes residual in residual structure, multi-skip connection and channel-attention module in SISR task. As the best performance record holder in 2018, it achieves better accuracy and visual improvement against the state-of-the-art in terms of PSNR and SSIM.

DPE: The deep photo enhancer (DPE) [1] learns image enhancement from a set of given photographers with unpaired settings. Inspired by CycleGAN, it works in a two-way GAN and makes some improvements on the stability and feature representation of GAN models. As a practical unsupervised method, it can be personalized to match individual users' preferences conveniently.

DRBN: A deep recursive band network (DRBN) is proposed to recover a linear band representation of an enhanced

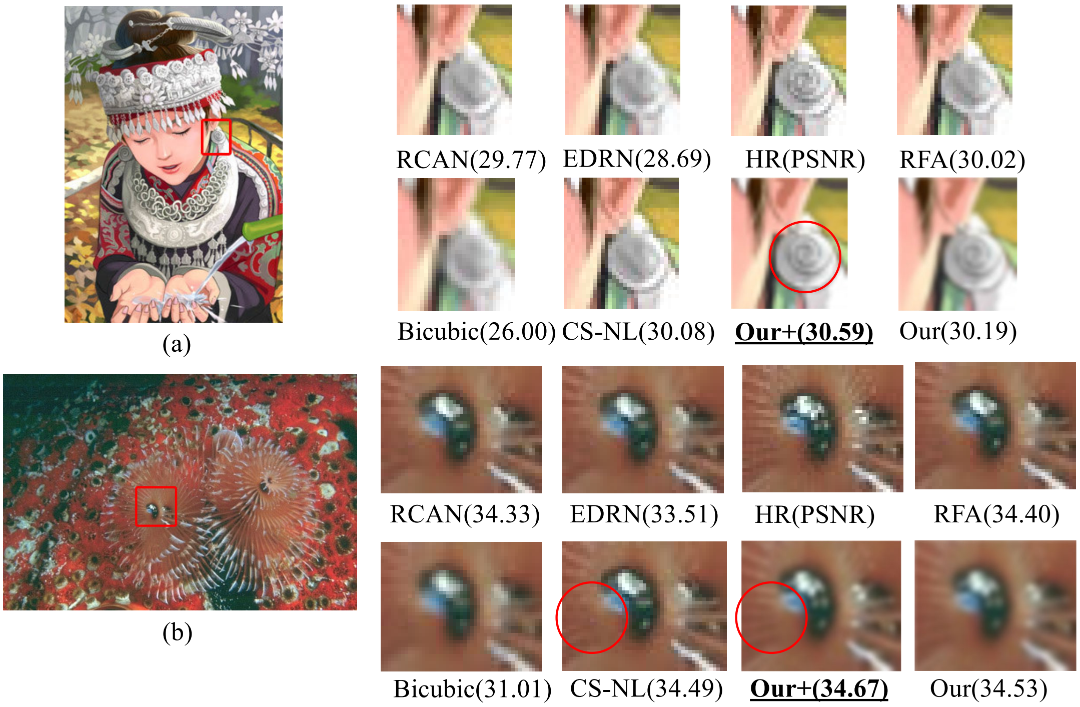


FIGURE 9. Qualitative comparison against the state-of-the-art on dataset SRT. Zoom in to see the details. The best result is highlighted with underline.

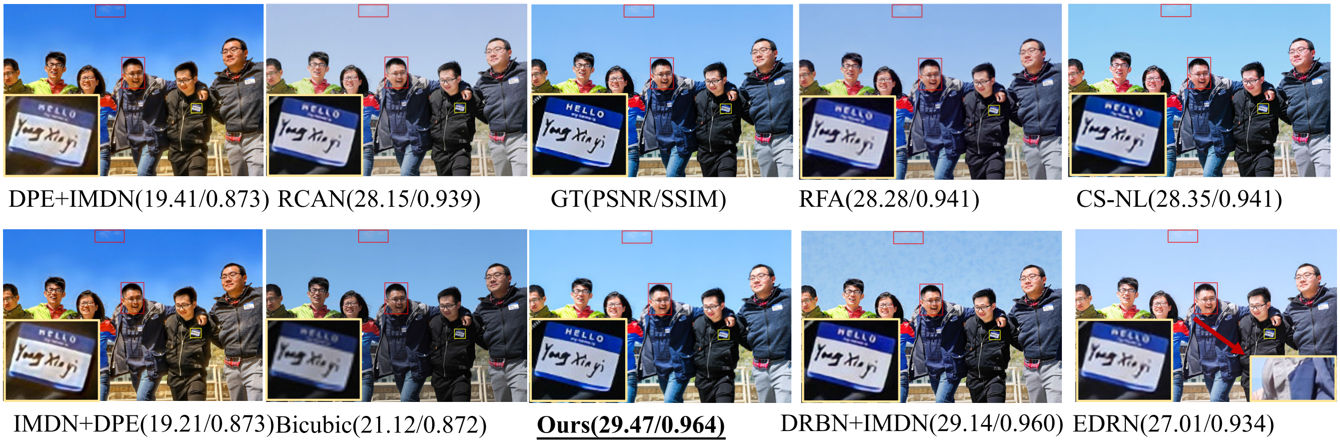
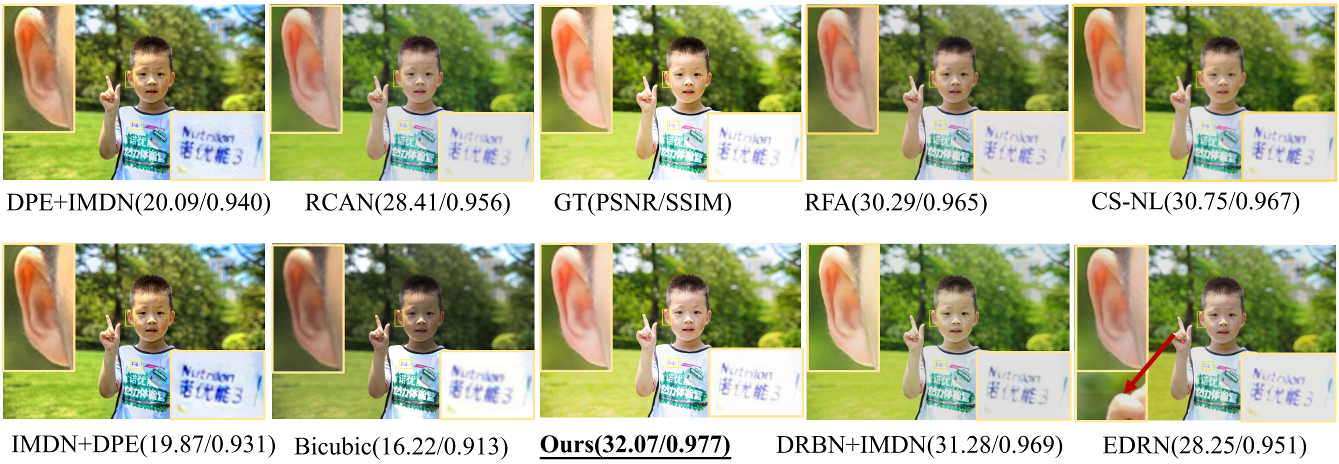


FIGURE 10. Qualitative comparison against the state-of-the-arts on dataset Alltuu2. Zoom in to see the details. The best result is highlighted with underline.

normal-light image with paired low/normal-light images, and then be improved via another learnable linear transformation based on a perceptual quality-driven adversarial learning with unpaired data [30].

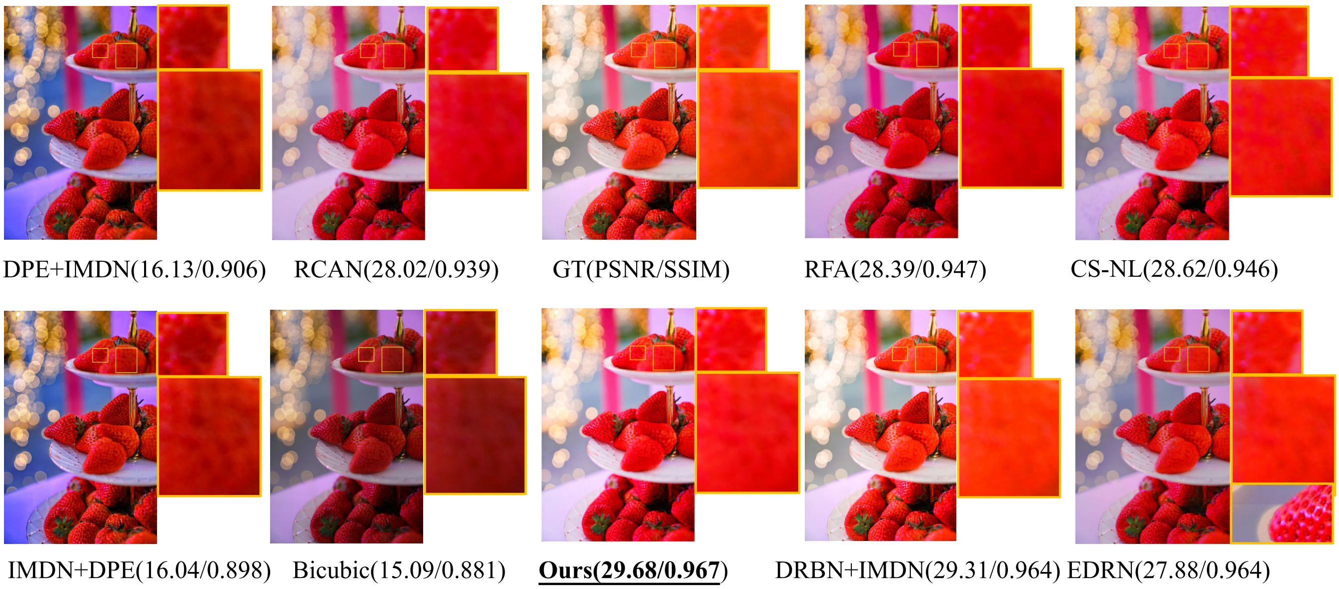
2) QUANTITATIVE COMPARISON

To confirm the effectiveness and robustness of our proposed model, we compare our method against the state-of-the-arts introduced in section IV-D1. To be fair, all the compared methods are trained on dataset SRT in SR task. We list the qualitative comparison with ×2 scaling factor and present their results in Table 2. The comparisons are organized into two groups based on different datasets. Since the well-trained methods on dataset SRT did not capture the perceptual change, they are unable to handle SR and PIE simultaneously. Therefore, in the joint task of SR and PIE, all the models are retrained on dataset Alltuu2 with the parameters declared in their papers. As in other SR methods, we also adopt self-ensemble strategy to further improve our framework and denote it as Deep SR-PIE+. For dataset Alltuu2, considering the joint-learning strategy, our Deep SR-PIE and Deep SR-PIE+ greatly surpass the others in the joint task. Without the consideration of perceptual recovery, it is observed that traditional SR methods (EDRN, CS-NL, RFA and RCAN) are capable of constructing the lost


DPE+IMDN(20.09/0.940) RCAN(28.41/0.956) GT(PSNR/SSIM) RFA(30.29/0.965) CS-NL(30.75/0.967)

IMDN+DPE(19.87/0.931) Bicubic(16.22/0.913) **Ours(32.07/0.977)** DRBN+IMDN(31.28/0.969) EDRN(28.25/0.951)

FIGURE 11. Qualitative comparison against the state-of-the-art on dataset Alltuu2. Zoom in to see the details. The best result is highlighted with underline.


DPE+IMDN(16.13/0.906) RCAN(28.02/0.939) GT(PSNR/SSIM) RFA(28.39/0.947) CS-NL(28.62/0.946)

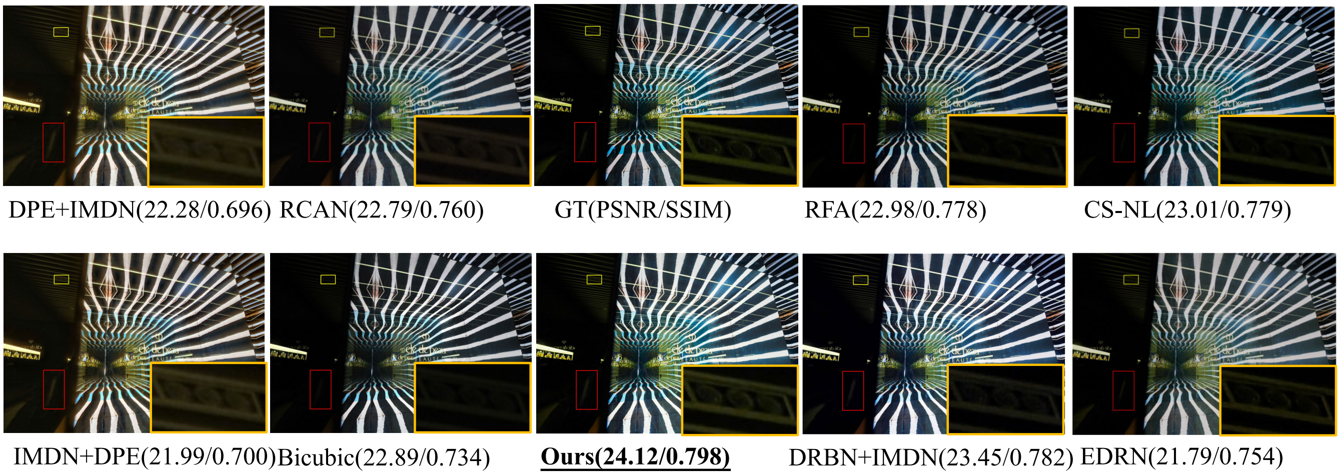
IMDN+DPE(16.04/0.898) Bicubic(15.09/0.881) **Ours(29.68/0.967)** DRBN+IMDN(29.31/0.964) EDRN(27.88/0.964)

FIGURE 12. Qualitative comparison against the state-of-the-art on benchmark dataset Alltuu2. Zoom in to see the details. The best result is highlighted with underline.

detail to some degree. Specially, DRBN + IMDN employs DRBN for light enhancement and then tackle SR with lightweight IMDN. Compared with the second best method (DRBN + IMDN), our method gains 2.48% and 0.65% performance on metrics PSNR and SSIM. Here, it is noting that DPE + IMDN and IMDN + DPE are only better than Bicubic. Limited by the ability of DPE in dark environment, these two models suffer from halo artifacts and amplified noise. Regarding dataset SRT, only our method is trained from the scratch while the others load pre-trained models on dataset DIV2K. From Table 2, on dataset Set14, Set15, BSD100 and Urban100, we find that our method all achieves best performance in terms of PSNR and SSIM. Therefore, we can safely come to the conclusion that our method produces comparable results even through it is not dedicated to SR task.

3) QUALITATIVE COMPARISON

To evaluate visual quality of the generated images, we place reconstructed results and some zoomed details in Fig. 9-Fig. 13. For dataset Alltuu2, the samples are randomly chosen with diverse properties including indoor and outdoor scenes. Comparing with other results, we notice three main observations of our method: 1). Our method is competent to recover more details and better contrast in all samples without obviously sacrificing over/under exposing parts. 2) It also produces vivid and natural color, making the reconstructed results more realistic. 3) It can eliminate noise and artifacts, leading to more visual pleasing results. Now, let's respectively take a closer at qualitative examples in Fig. 10-Fig. 13. In Fig. 10, it can seen DRBN + IMDB is liable to produce artifacts in blue sky (the area marked by red box) while RCAN always renders slightly overexposed


DPE+IMDN(22.28/0.696) RCAN(22.79/0.760) GT(PSNR/SSIM) RFA(22.98/0.778) CS-NL(23.01/0.779)

IMDN+DPE(21.99/0.700) Bicubic(22.89/0.734) **<u>Ours(24.12/0.798)</u>** DRBN+IMDN(23.45/0.782) EDRN(21.79/0.754)

FIGURE 13. Qualitative comparison against the state-of-the-art on benchmark dataset Alltuu2. Zoom in to see the details. The best result is highlighted with underline.

results. Fig. 11 shows a boy stands in a park who is posing for taking picture. As can be seen, our method shows clear characters in the T-skirt with natural light condition. In Fig. 12, several methods fail to locate the rotten part of strawberries (CS-NL and EDRN), and could not provide realistic color rendition of light spots on the top left corner (DPE + IMDN,IMDN + DPE, Bicubic, CS-NL, RFA). In Fig. 13, under extreme low-light condition, the majority of developed methods could not perceive objects clearly hidden in the left dark area. Surprisingly, the result processed by Bicubic achieves relatively higher PSNR but lower SSIM. The main reason can be illustrated as the difference between original input image and enhanced image is slighter than other samples. Though it is true that Bicubic achieves better PSNR in this case, its result looks worse from the perceptive of visual perception. Inherited from tone mapping for some images, the methods related to DPE (DPE + IMDN and IMDN + DPE) have trouble in producing clear and clean results. Besides, one can also see that all samples yielded by EDRN have horizontal patch borders that look like crease mark where the arrows are pointing. For dataset SRT, we depict two sets of visual results processed by the comparative methods in Fig. 9. Compared with the others, the images reconstructed by our methods (Deep SR-PIE and Deep SR-PIE+) yield compelling visual effect. To be specific, In Fig. 9(a), only Deep SR-PIE and Deep SR-PIEd+ display clear texture of silver eardrop. Also for the anemones in Fig. 9(b), Deep SR-PIE (Ours) and Deep SR-PIEd (Ours+) can help us to count the number of tentacles while the others cannot. As a conclusion, the above comparison on datasets Alltuu2 and SRT can demonstrate the effectiveness and robustness of our method both in single SR task and joint SR-PIE task.

4) EVALUATION OF TIME EFFICIENCY

In this section, we study time efficiency of our proposed framework. We compare our model with EDRN, RFA, CS-NL, RCAN and DEP + IMDB on dataset Alltuu2. Notably, all comparisons are evaluated on the same machine. As seen in Table 2, with distillation model and selective fused parts, DEP + IMDB and Deep SR-PIE (Ours) are compatible with low computing power devices. Unsurprisingly, by introducing more elaborate and complex modules, both the size and the running time of the comparative models increase by a large margin. Both on model size and inference time, our method demonstrates impressive and feasible capability.

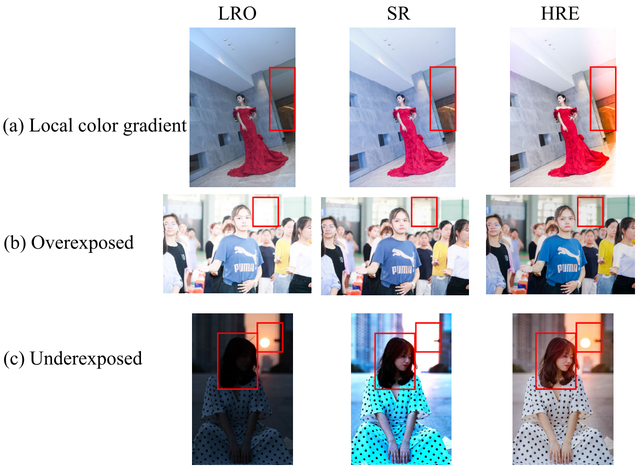


FIGURE 14. Failure Cases. (a) The condition of local color gradient. (b) Overexposed condition. (c) Underexposed condition.

5) FAILURE CASES STUDY

Though our proposed framework works well on the majority of testing images, we list several failure cases in Fig. 14. In Fig. 14(a), except for color adjustment and contrast balance, retouchers might use local color gradient to give the whole scene a more natural feeling that cannot be learned by our deep model. In our testing example, an additional

TABLE 2. Quantitative comparison of different approaches for SR-PIE on datasets: Alltuu2, Set14, BSD100, Urban100 and Set 5. The SR factor is 2. The best results are highlighted in bold. Params(M): The number of parameters (unit:million). Time(s): The inference time (unit:second). '/' denotes the result is not available.

Methods	Alltuu2				Set14		BSD100		Urban100		Set5	
	PSNR	SSIM	Params(M)	Time(s)	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
Bicubic	18.0148	0.8681	/	/	30.24	0.8688	29.56	0.8431	26.88	0.8403	33.66	0.9299
RFA	28.6200	0.9434	11.14	5.214	34.16	0.9220	32.41	0.9026	33.33	0.9389	38.26	0.9615
EDRN	27.0256	0.9393	58.49	4.43	33.65	0.9185	32.29	0.9010	32.35	0.9307	38.13	0.9609
CS-NL	28.9105	0.9573	3	6.635	34.12	0.9223	32.40	0.9024	33.25	0.9386	38.28	0.9616
RCAN	28.0124	0.9398	15.59	7.125	34.11	0.9216	32.41	0.9026	33.34	0.9284	38.27	0.9614
DPE+IMDN	19.8399	0.8892	/	/	/	/	/	/	/	/	/	/
IMDN+DPE	19.5657	0.8832	/	/	/	/	/	/	/	/	/	/
DRBN+IMDN	29.1156	0.9587	0.77	**0.43**	33.65	0.9173	32.23	0.9103	32.20	0.9286	38.09	0.9609
Ours	29.8379	0.9649	**0.76**	0.46	34.20	0.9228	32.46	0.9030	33.39	0.9392	38.31	0.9619
Ours+	**29.8563**	**0.9684**	0.77	1.85	**34.21**	**0.9230**	**32.48**	**0.9031**	**33.40**	**0.9395**	**38.33**	**0.9620**

lighting is introduced, which alters the image tone irrevocably. In Fig. 14(b), when too much light is allowed during exposure, the image brighter than it should be is often considered overexposed. Seeing that too much details are lost in this mode, our framework is incapable of restoring the imponderable pixels. As we can seen, the detail of window frames in the background cannot be reconstructed as rich as the one in ground-truth. In Fig. 14(c), since severely-underexposed image is usually imperceptible and its enhancement is highly nonlinear and subjective, it is infeasible to fully recover the lost details. In our case, our model fails to reconstruct the girl's facial expression under extreme conditions, since the region is almost black without any trace in input image.

V. CONCLUSION AND FUTURE WORK

In this paper, we propose the Deep SR-PIE, a joint learning framework for SR-PIE task. To boost the capacity of precisely predicting lost high-frequency details, the original input is decomposed and adopted for different branches. For branch MBFNet, Multi-scale Backward Fusion Network (MBFNet) shoulders the responsibility for SR task by fusing hierarchical deep features. The branch of Perceptual Enhancement Network (PENet) aims to learn the perceptual mapping from LRO to HRE, which could assist our model in recovering the lost of color, tone, contrast and so on. For branch Dual-path Upsampling Net (DUNet), it provides an informative upsampling feature map via shared bypath convolutions, which is conducive to capturing lost details to some extent. Besides, the comprehensive experimental results have demonstrated that our proposed Deep SR-PIE delivers comparative performance against the state-of-the-arts on four datasets.

Our further work is to adaptively solve SR-PIE problem of arbitrary scale factor with a single model. Another direction is to address the nearly black and white region by virtue of semantic analysis and image content generation.

REFERENCES

[1] Y.-S. Chen, Y.-C. Wang, M.-H. Kao, and Y.-Y. Chuang, "Deep photo enhancer: Unpaired learning for image enhancement from photographs with GANs," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 6306–6314.

[2] C. Dong, C. C. Loy, K. He, and X. Tang, "Learning a deep convolutional network for image super-resolution," in *Proc. Eur. Conf. Comput. Vis.* New York, NY, USA: Springer, 2014, pp. 184–199.

[3] R. Wang, Q. Zhang, C.-W. Fu, X. Shen, W.-S. Zheng, and J. Jia, "Underexposed photo enhancement using deep illumination estimation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 6849–6857.

[4] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[5] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 4700–4708.

[6] 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.

[7] J. Kim, J. K. Lee, and K. M. Lee, "Accurate image super-resolution using very deep convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 1646–1654.

[8] K. Zhang, W. Zuo, S. Gu, and L. Zhang, "Learning deep CNN denoiser prior for image restoration," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 3929–3938.

[9] C. Ledig, L. Theis, F. Huszar, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, Z. Wang, and W. Shi, "Photo-realistic single image super-resolution using a generative adversarial network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 4681–4690.

[10] T. Tong, G. Li, X. Liu, and Q. Gao, "Image super-resolution using dense skip connections," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 4799–4807.

[11] J. Liu, W. Zhang, Y. Tang, J. Tang, and G. Wu, "Residual feature aggregation network for image super-resolution," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 2359–2368.

[12] Y. Tai, J. Yang, X. Liu, and C. Xu, "MemNet: A persistent memory network for image restoration," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 4539–4547.

[13] Y. Zhang, Y. Tian, Y. Kong, B. Zhong, and Y. Fu, "Residual dense network for image super-resolution," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 2472–2481.

[14] N. Ahn, B. Kang, and K.-A. Sohn, "Fast, accurate, and lightweight super-resolution with cascading residual network," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 252–268.

[15] M. Haris, G. Shakhnarovich, and N. Ukita, "Deep back-projection networks for super-resolution," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 1664–1673.

[16] T. Dai, J. Cai, Y. Zhang, S.-T. Xia, and L. Zhang, ''Second-order attention network for single image super-resolution,'' in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 11065–11074.

[17] Y. Mei, Y. Fan, Y. Zhou, L. Huang, T. S. Huang, and H. Shi, ''Image super-resolution with cross-scale non-local attention and exhaustive self-exemplars mining,'' in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 5690–5699.

[18] J. Cai, S. Gu, and L. Zhang, ''Learning a deep single image contrast enhancer from multi-exposure images,'' *IEEE Trans. Image Process.*, vol. 27, no. 4, pp. 2049–2062, Apr. 2018.

[19] J. Yan, S. Lin, S. B. Kang, and X. Tang, ''A Learning-to-Rank approach for image color enhancement,'' in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 2987–2994.

[20] Z. Yan, H. Zhang, B. Wang, S. Paris, and Y. Yu, ''Automatic photo adjustment using deep neural networks,'' *ACM Trans. Graph.*, vol. 35, no. 2, pp. 1–15, May 2016.

[21] 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, pp. 1–12, Jul. 2017.

[22] Y. Hu, H. He, C. Xu, B. Wang, and S. Lin, ''Exposure: A white-box photo post-processing framework,'' *ACM Trans. Graph.*, vol. 37, no. 2, pp. 1–17, Jul. 2018.

[23] J. Park, J.-Y. Lee, D. Yoo, and I. S. Kweon, ''Distort- and-recover: Color enhancement using deep reinforcement learning,'' in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 5928–5936.

[24] 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.

[25] M. S. M. Sajjadi, B. Scholkopf, and M. Hirsch, ''EnhanceNet: Single image super-resolution through automated texture synthesis,'' in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 4491–4500.

[26] 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.

[27] 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.

[28] W. Wang, Q. Lai, H. Fu, J. Shen, H. Ling, and R. Yang, ''Salient object detection in the deep learning era: An in-depth survey,'' 2019, *arXiv:1904.09146*. [Online]. Available: http://arxiv.org/abs/1904.09146

[29] Y. Jiang, X. Gong, D. Liu, Y. Cheng, C. Fang, X. Shen, J. Yang, P. Zhou, and Z. Wang, ''EnlightenGAN: Deep light enhancement without paired supervision,'' 2019, *arXiv:1906.06972*. [Online]. Available: http://arxiv.org/abs/1906.06972

[30] W. Yang, S. Wang, Y. Fang, Y. Wang, and J. Liu, ''From fidelity to perceptual quality: A semi-supervised approach for low-light image enhancement,'' in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 3063–3072.

[31] E. Schwartz, R. Giryes, and A. M. Bronstein, ''DeepISP: Toward learning an End-to-End image processing pipeline,'' *IEEE Trans. Image Process.*, vol. 28, no. 2, pp. 912–923, Feb. 2019.

[32] X. Xu, Y. Ma, and W. Sun, ''Towards real scene super-resolution with raw images,'' in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 1723–1731.

[33] K. Mei, J. Li, J. Zhang, H. Wu, J. Li, and R. Huang, ''HighEr-resolution network for image demosaicing and enhancing,'' 2019, *arXiv:1911.08098*. [Online]. Available: http://arxiv.org/abs/1911.08098

[34] 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.

[35] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. Alemi, ''Inception-v4, inception-ResNet and the impact of residual connections on learning,'' 2016, *arXiv:1602.07261*. [Online]. Available: http://arxiv.org/abs/1602.07261

[36] Z. Hui, X. Gao, Y. Yang, and X. Wang, ''Lightweight image super-resolution with information multi-distillation network,'' in *Proc. 27th ACM Int. Conf. Multimedia*, Oct. 2019, pp. 2024–2032.

[37] F. Yu, V. Koltun, and T. Funkhouser, ''Dilated residual networks,'' in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 472–480.

[38] M. Afifi, B. Price, S. Cohen, and M. S. Brown, ''When color constancy goes wrong: Correcting improperly white-balanced images,'' in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 1535–1544.

[39] Z. Lu, C. Liu, and X. Zhong, ''Bezier curve-based saturation-aided optimal brightness adjustment for dark image clearness enhancement with image fusion,'' *Signal, Image Video Process.*, vol. 14, pp. 1625–1633, 2020.

[40] W. Shi, J. Caballero, F. Huszar, J. Totz, A. P. Aitken, R. Bishop, D. Rueckert, and Z. Wang, ''Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network,'' in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 1874–1883.

[41] J. Johnson, A. Alahi, and L. Fei-Fei, ''Perceptual losses for real-time style transfer and super-resolution,'' in *Proc. Eur. Conf. Comput. Vis.* New York, NY, USA: Springer, 2016, pp. 694–711.

[42] H. Zhao, O. Gallo, I. Frosio, and J. Kautz, ''Loss functions for image restoration with neural networks,'' *IEEE Trans. Comput. Imag.*, vol. 3, no. 1, pp. 47–57, Mar. 2017.

[43] E. Agustsson and R. Timofte, ''NTIRE 2017 challenge on single image super-resolution: Dataset and study,'' in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jul. 2017, pp. 126–135.

[44] Z. Wang, J. Chen, and S. C. H. Hoi, ''Deep learning for image super-resolution: A survey,'' *IEEE Trans. Pattern Anal. Mach. Intell.*, early access, Mar. 23, 2020, doi: 10.1109/TPAMI.2020.2982166.

[45] D. Martin, C. Fowlkes, D. Tal, and J. Malik, ''A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics,'' in *Proc. 8th IEEE Int. Conf. Comput. Vis. (ICCV)*, vol. 2, Jul. 2001, pp. 416–423.

[46] R. Zeyde, M. Elad, and M. Protter, ''On single image scale-up using sparse-representations,'' in *Proc. Int. Conf. Curves Surf.* New York, NY, USA: Springer, 2010, pp. 711–730.

[47] J.-B. Huang, A. Singh, and N. Ahuja, ''Single image super-resolution from transformed self-exemplars,'' in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 5197–5206.

[48] C. Guo, C. Li, J. Guo, C. C. Loy, J. Hou, S. Kwong, and R. Cong, ''Zero-reference deep curve estimation for low-light image enhancement,'' in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 1780–1789.

[49] W. Sun and Z. Chen, ''Learned image downscaling for upscaling using content adaptive resampler,'' *IEEE Trans. Image Process.*, vol. 29, pp. 4027–4040, 2020.

[50] G. Cheng, A. Matsune, Q. Li, L. Zhu, H. Zang, and S. Zhan, ''Encoder-decoder residual network for real super-resolution,'' in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2019, pp. 156–162.

[51] Y. Zhang, K. Li, K. Li, L. Wang, B. Zhong, and Y. Fu, ''Image super-resolution using very deep residual channel attention networks,'' in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 286–301.

YIFEI XU received the B.S. degree in computer science and technology from the South China University of Technology (SCUT), Guangzhou, China, in 2011, and the Ph.D. degree from the School of Computer Science, Zhejiang University, in 2017. In 2017, he was with the School of Software, Xi'an Jiaotong University, as an Assistant Professor. His current research interests include deep learning, image enhancement, and image semantic segmentation.

NUO ZHANG received the B.S. degree in Internet of Things from Hainan University, Hainan, China, in 2014. She is currently pursuing the M.Eng. degree with the School of Software, Xi'an Jiaotong University. Her research interests include deep learning and computer vision.

LI LI received the B.S. degree from the South China University of Technology, in 2011, and the M.S. degree from Zhejiang University, in 2014. He is currently working as a Researcher at the ZhiTu Lab in Alltuu. His main research interests include computer vision, image processing, and deep learning.

GENAN SANG received the B.S. degree from Nanjing Tech University, in 2012. He received the M.S. degree from Hangzhou Normal University, in 2019. He is currently a Researcher at the ZhiTu Lab in Alltuu. His research interests include image and video editing, computer vision, and computational photography.

YUEWAN ZHANG received the B.S. degree in computer science from Zhejiang Gongshang University, in 2019. She is currently pursuing the master's degree in software engineering with Xi'an Jiaotong University. Her research interests include computer vision and deep learning.

ZHENGYANG WANG received the B.S. degree in chemical engineering and technological from the Changchun University of Science and Technology, Jilin, China, in 2018. He is currently pursuing the degree in software engineering with Xi'an Jiaotong University. His professional direction is cloud computing.

PINGPING WEI received the B.S. degree in computer science and technology from Tianjin University, China, in 2012, and the M.S. degree from The University of Sydney, Australia. She is currently working with Xi'an Jiaotong University.