IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.

Digital Object Identifier 10.1109/ACCESS.2017.DOI

Guided Dual Networks for Single Image Super-Resolution

WENHUI CHEN[1], CHUANGCHUANG LIU[1], YITONG YAN[1], LONGCUN JIN[1] (Member, IEEE), XIANFANG SUN[2], XINYI PENG[1]
[1]School of Software Engineering, South China University of Technology, Guangzhou, China
[2]School of Computer Science and Informatics Cardiff University, UK

Corresponding author: Longcun Jin (e-mail: lcjin@scut.edu.cn).

This work was supported in part by the Pearl River Technology Nova Project under Grant 201710010020, the Fundamental Research Funds for the Central Universities under Grant 2019MS086, and the National Science Foundation of China under Grant 61300135.

ABSTRACT The PSNR-oriented super-resolution (SR) methods pursue high reconstruction accuracy, but tend to produce over-smoothed results and lose plenty of high-frequency details. The GAN-based SR methods aim to generate more photo-realistic images, but the hallucinatory details are often accompanied with unsatisfying artifacts and noise. To address these problems, we propose a guided dual super-resolution network (GDSR), which exploits the advantages of both the PSNR-oriented and the GAN-based methods to achieve a good trade-off between reconstruction accuracy and perceptual quality. Specifically, our network contains two branches, where one is trained to extract global information and the other to focus on detail information. In this way, our network simultaneously generates SR images with high accuracy and satisfactory visual quality. To obtain more high-frequency features, we use the global features extracted from the low-frequency branch to guide the training of the high-frequency branch. Besides, our method utilizes a mask network to adaptively recover the final super-resolved image. Extensive experiments on several standard benchmarks show that our proposed method achieves better performance compared with state-of-the-art methods. The source code and the results of our GDSR are available at https://github.com/wenchen4321/GDSR.

INDEX TERMS Convolutional neural network, Dual Network, Generative adversarial network, Single Image Super-Resolution.

I. INTRODUCTION

AIMING to recover a high-resolution (HR) image from a single given low-resolution (LR) image, single image super-resolution (SISR) has received critical attention in computer vision researches. SISR can be used in various fields, such as security and surveillance [1], [2], medical imaging [3], [4], remote sensing image [5], and object recognition [6]–[9]. However, because there exist multiple solutions for the same low-resolution image, SISR is an ill-posed inverse problem. Most SR methods learn the mapping between LR and HR images for generating high-quality super-resolved images.

In recent years, deep convolutional neural network (CNN) based methods [10], [11] have consistently achieved significant improvements over traditional methods in reconstructing high-quality SR images. Various network architectures are proposed to improve the SR performance, commonly taking the Peak Signal-to-Noise Ratio (PSNR) and/or the Structural Similarity Index (SSIM) [12] as measurements and evaluation indexes. These methods assume that higher PSNR value implies better quality and less distortion. They usually adopt the optimization function of minimizing the mean

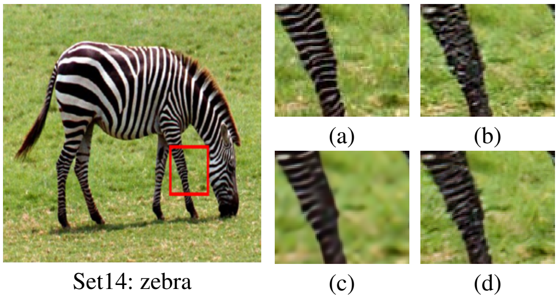


FIGURE 1. The 4× SR visual comparisons of 'zebra' in Set14. (a) the HR image; (b) GAN-based method (ESRGAN); (c) PSNR-oriented method (RCAN); (d) Our network (GDSR).

squared error (MSE) between the recovered SR image and the ground truth to maximize PSNR. However, they lack the ability to capture high-frequency features. The high PSNR estimates are typically over-smoothed and conflict with human visual perceptual observation. As shown in Fig. 1, the PSNR-oriented method RCAN [13] has high reconstruction accuracy but generates over-smoothed edges.

To improve the visual quality of SR images, several perceptual-driven methods have been proposed to produce visually satisfying results. Generative adversarial networks (GANs) [14] have been applied in SR because of its capability to generate realistic images. Despite their great success, most GAN-based SR methods pay too much attention to the high-frequency information in the super-resolved images. Although the generated images can recover more details, sometimes they are noisy and bring unpleasing artifacts. As shown in Fig. 1, the image generated from the GAN-based method ESRGAN [15] is noisy. The process of SR is sometimes treated as a pre-processing step for other high-level computer vision tasks such as object recognition and image classification. The noise and artifacts generated by GAN-based SR methods would be detrimental to high-level computer vision tasks.

In general, the existing dual branch SISR methods, such as DualCNN [16] and Dual-way SR [17], learn the global information and the details from two branches with different network structures. They are still PSNR-oriented methods and produce over-smoothed results as they use a single or the same loss function for both branches. We will design a model with two branches of the same network structure. However, we supervise two branches to learn different information using different loss functions, where one branch is related to the PSNR-oriented method and the other is related to the GAN-based method. In this way, better performance than PSNR-oriented dual-branch methods can be achieved.

The PSNR-oriented methods produce high accuracy SR results with over-smoothed edges, while the GAN-based methods generate SR images with better perceptual quality but sometimes with noise and artifacts. Either of them could not balance the accuracy and the perceptual quality. Inspired by the dual skipping network [18], which is used for coarse-to-fine object categorization, we propose a guided dual SR network (GDSR) to achieve a good trade-off between perceptual quality and reconstruction accuracy. Our method reconstructs images with high visual quality and less deformed textures in a left-right asymmetric network. Specifically, our network has two branches: the high-frequency branch (HFB) and the low-frequency branch (LFB). Trained with the GAN adversarial loss, the HFB aims to extract high-frequency features and make the SR image contain more detail information. The LFB is trained with the MSE loss to extract global information. Similar to the dual skipping network, we adopt a top-down global guidance mechanism to guide the HFB. In brief, the guidance feeds the high-level global information from the LFB to the corresponding lower-level feature processing modules of the HFB. Furthermore, we use a mask network to produce an attention mask for weighting the outputs of the LFB and the HFB, adaptively recovering the final super-resolved image. As shown in Fig. 1, the SR image generated by our GDSR is more accurate and faithful to the ground truth. It achieves better visual SR result compared with state-of-the-art methods.

The contributions of this paper are summarized as follows:

- We propose a left-right asymmetric super-resolution network by integrating GAN-based and PNSR-oriented methods to improve the SR image quality. Our GDSR network can generate SR images with higher perceptual quality and less distortion.
- We employ the top-down global guidance to deliver the high-level global features from the low-frequency branch to the high-frequency branch for generating detail information.
- Extensive experiments show that our approach achieves state-of-the-art performance on several benchmarks, demonstrating the effectiveness of our network.

II. RELATED WORK

A. SINGLE IMAGE SUPER-RESOLUTION

Since deep learning algorithms have shown superior performance, we mainly focus on deep learning algorithms for the SISR problem.

SRCNN [19] is the first successful attempt towards super-resolution using only three convolutional layers. This effort can be considered as the pioneering work in the CNN-based SR field and inspires numerous later works. Replacing the bicubic upsampling operation with an efficient sub-pixel convolution, FSRCNN [20] and ESPCN [21] improve the speed and image quality, which achieve real-time performance. Various advanced upsampling structures have been proposed in recent years, such as deconvolutional layer [22], [23] and EUSR [24]. VDSR [25] and DRCN [26] increase the network depth and achieve better performance, supporting the argument that deeper networks can provide better contextualization. EDSR [27] introduces a very deep and wide network by modifying the ResNet [28] architecture. LapSRN [29] employs a pyramidal framework to progressively predict the residual images up to a factor of 8×. ZSSR [30] uses an unsupervised method to learn the mapping between LR and HR images. SRMDNF [31] tackles multiple degradation problems in a single network by treating degradation maps of images as inputs. Densely connected networks have been proposed to improve SR performance. RDN [32] combines residual skip connections with dense connections, showing good resilience against the degradation process and recovering enhanced SR images. RCAN [13] is a recently proposed deep ResNet network with the channel attention mechanism and achieves state-of-the-art PSNR performance.

The aforementioned methods mainly use MSE loss as the optimization function to obtain high PSNR and SSIM values. However, these PSNR-oriented methods usually generate heavily over-smoothed edges. The generated images lose various high-frequency details and have bad perceptual quality.

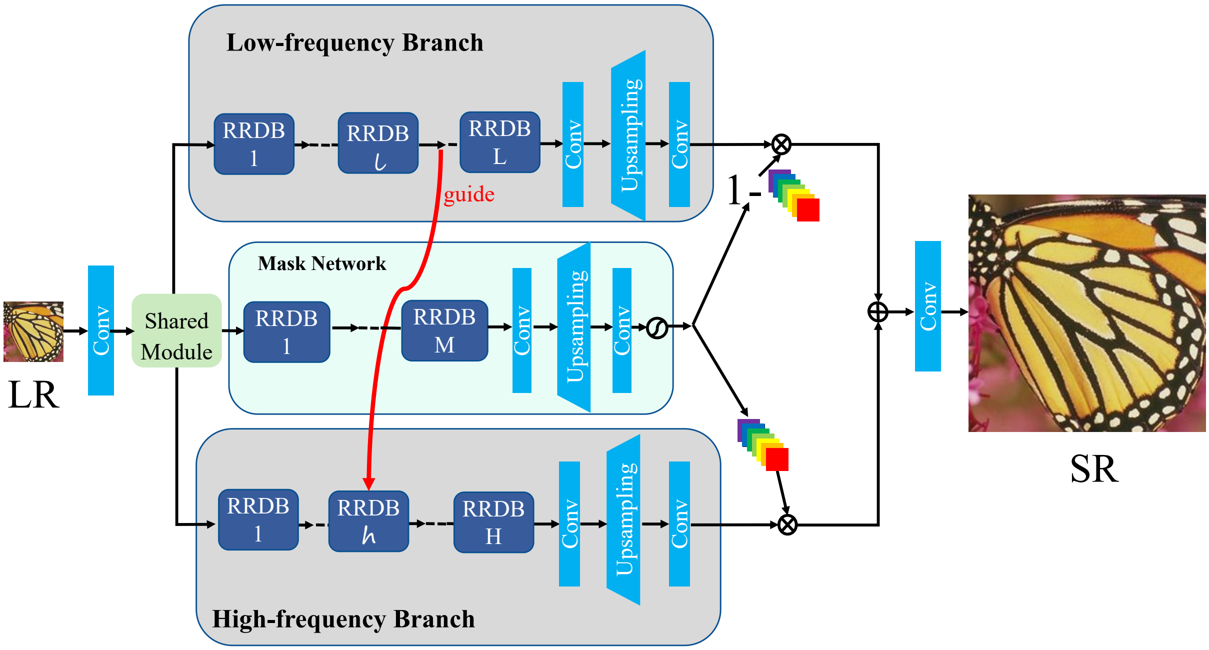


FIGURE 2. The overall architecture of our proposed method. It contains three subnets: the low-frequency branch (LFB), the mask network and the high-frequency branch (HFB), respectively. They share the same shallow feature extraction module, namely the shared module (SM). Each subnet is stacked with the basic blocks: RRDBs. The top-down guide delivers the global information from a high abstraction level of the LFB to a lower abstraction level of the HFB. The mask network generates an attention mask to combine the feature maps from the LFB and the HFB right before the last convolution layer of reconstruction.

To improve the visual quality of SR results, perceptual-driven approaches have been proposed. SRGAN [33] firstly introduces the GAN framework into the SR problem and produces visually pleasing results. SRGAN combines a perceptual loss and an adversarial loss to improve the reality of the generated images. But visually implausible artifacts can still be found in some generated images. EnhanceNet [34] combines a pixel-wise loss in the image space, a perceptual loss in the feature space, an adversarial loss, and a texture matching loss [35] to produce more realistic and better perceptual-quality outputs. Built upon SRGAN, ESRGAN [15] removes batch normalization layers and introduces a basic and effective block: Residual-in-Residual Dense Block (RRDB). Moreover, ESRGAN also employs an enhanced discriminator called Relativistic average GAN (RaGAN) [36]. Noteworthily, ESRGAN won the first place in the 2018 PIRM Challenge on Perceptual Image Super-Resolution [37], which evaluates the image perceptual quality using the perceptual index (PI). The SR models trained with the MSE loss tend to produce over-smoothed results while that trained in an adversarial manner generate realistic details but bring some unpleasant noise. By simply utilizing linear network interpolation of the results generated from PNSR-oriented and GAN-based models, DNI [38] balances the MSE and GAN effects of SR results. But the interpolation parameter α is selected manually. It is too costly to generate continuous interpolation results from interpolating the PSNR-oriented model and the GAN-based method with parameter α in $[0, 1]$. EEGAN [39] proposes a GAN-based edge-enhancement network that has two subnetworks: a GAN-based ultradense subnetwork and a CNN-based edge-enhancement subnetwork. However, EEGAN is specifically designed for satellite image SR reconstruction, where the CNN-based edge-enhancement subnetwork is for extracting the special features of the edges from satellite images. RankSRGAN [40] introduces a ranker to optimize the perceptual metric directly, which only pursues lower PI value and brings blurring artifacts to the hallucinated details.

Some SISR methods use two branches to capture more information to achieve better performance. DualCNN [16] is a PSNR-oriented network, which uses the different numbers of convolution layers to extract the structure information and the details. SRDPN [41] replaces the residual blocks of EDSR with DPN [42] blocks to achieve improved performance. DSRN [43] introduces a dual-state recurrent network to incorporate information from both the LR and the HR spaces. Dual-way SR [17] exploits a complex network EDSR as its complex branch and the bicubic interpolation as its plain branch to capture the global and the detail information. The above dual-path methods design different network structures for different branches to capture more information, but all the branches of these methods are trained with the same loss function. They are still PSNR-oriented methods and cannot solve the problem of generating over-smoothed results. We use dual branches to learn different information based on different loss functions, rather than using different network structures. Our network leverages the advantages of both the PSNR-oriented and the GAN-based methods to supervise the network and capture the high-frequency and the low-frequency features. The proposed training strategies facilitate reconstructing accurate and realistic super-resolution images.

B. DUAL SKIPPING NETWORK

The study on hemispheric specialization shows that visual analysis takes place in a predominately and default coarse-to-fine sequence. Instead of processing spatial frequency information equally, the recent biological experiments reveal that the left hemisphere (LH) and the right hemisphere (RH) are predominantly involved in the high and low spatial frequency processing, respectively. Inspired by the research of the primate visual cortex, the dual skipping network [18] shows promising results on coarse-to-fine object categorization. The dual skipping network is a left-right asymmetric layer skippable network which has two branches referring to LH and RH, respectively. One branch is used for fine-grained level classification which simulates the LH mechanism of processing spatial high-frequency information. The other branch is used for coarse-level classification which simulates the RH mechanism of processing spatial low-frequency information. So the dual skipping network can simultaneously work on coarse and fine-grained classification tasks. Moreover, motivated by a similar mechanism in the brain, the dual skipping network introduces a "Guide" referring to top-down facilitation of recognition. The guide feeds the high-level information from the coarse branch to relatively lower-level visual processing modules of the fine-grained branch. In our network, inspired by the study on hemispheric specialization and dual skipping network, we design the high-frequency branch to simulate the LH processing mechanism, and the asymmetric low-frequency branch to process the RH mechanism. Moreover, we also design a mask network to simulate the function of the cerebellum, which is involved in balance and motor control.

III. PROPOSED METHOD

This section introduces the proposed method in detail. Our GDSR network aims to improve the perceptual quality and reconstruction accuracy of SR images via a left-right asymmetric network architecture. As shown in Fig. 2, the GDSR consists of three key components: 1) a left-right asymmetric SR network, 2) a global guidance mechanism, and 3) a mask network. The left-right asymmetric network architecture mainly consists of two branches: the high-frequency branch (HFB) and the low-frequency branch (LFB). The guidance delivers the global feature maps from the LFB to the low-level module of the HFB, helping the HFB generate more high-frequency details. The mask network adaptively reconstructs the final output from the LFB and the HFB to improve the perceptual quality and reconstruction accuracy of the SR image. We first describe our left-right asymmetric SR network architecture, then detail our guidance mechanism and mask network in the later subsections.

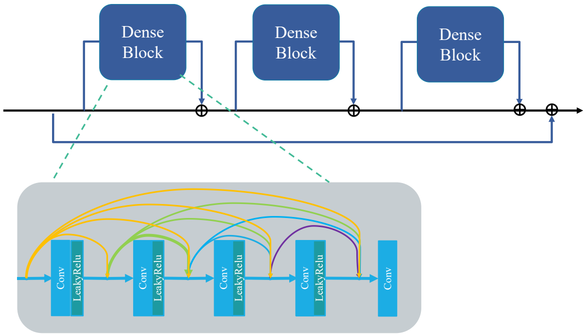


FIGURE 3. The basic block in our network: Residual-in-Residual Dense Block (RRDB). The RRDB has a global residual connection, while each dense block has a local residual connection and dense connections inside the block.

A. LEFT-RIGHT ASYMMETRIC NETWORK ARCHITECTURE

Our left-right asymmetric SR network aims to achieve a better trade-off between reconstruction accuracy and perceptual quality, which contains two complementary branches. The HFB is used to recover detail information, while the LFB is for reconstructing global information. ESRGAN [15] introduces a novel effective basic block: Residual-in-Residual Dense Block (RRDB), as depicted in Fig. 3, to generate high-quality images. The excellent experiment results prove the strong ability of RRDB to extract multi-level feature information, and we also utilize the RRDB as our basic block.

1) Shared Module

Generally, the shallow parts of the three subnets always extract the shallow features, such as edges and corners. Hence, we design a shared module (SM) for these three subnets at the head of our framework. We first use a convolution layer to process the same LR input image of the subnets, attaining the feature map F_0 of the input:

$$F_0 = H_{LR}(I_{LR}), \tag{1}$$

where H_{LR} represents convolution operation. Then we use S RRDBs in the shared module to obtain shallow features from the input feature map F_0, so we can have:

$$F_{SF} = H_{SM}(F_0), \tag{2}$$

where H_{SM} denotes our shared module. The SM is shared by the low-frequency branch, the high-frequency branch and the mask network, which can extract feature maps efficiently and reduce the parameters.

2) Low-Frequency Branch

To reconstruct relatively high-accuracy SR images, we use a low-frequency branch (LFB) to extract the global features. The LFB is trained by the MSE loss, including deep level feature extraction module, upsampling module and reconstruction module with a feed-forward pipeline. We adopt L RRDBs in the deep feature extraction module to obtain more global feature maps, while the upsampling module upscales the feature maps and the reconstruction module outputs the super-resolved feature maps by a convolution layer.

3) High-Frequency Branch

To produce more photo-realistic SR images, we utilize the HFB to generate more high-level feature maps. The HFB is trained by the GAN framework having a generator and a discriminator. The network structure of the generator is similar to the LFB, where H RRDBs are stacked, followed by a convolution layer, an upsampling layer and another convolution layer for reconstruction. The discriminator is a classification network to distinguish the real HR image and the artificially super-resolved image. Similar to ESRGAN [15], we apply the Relativistic average Discriminator (RaD) [36] as our HFB discriminator. The probability output of RaD being closer to 1 means the real image x_r being more realistic than the fake one x_f. The loss function of the discriminator is defined as follows:

$$\begin{aligned} L_D^{Ra} = & -\mathbb{E}_{x_r}[\log(D_{Ra}(x_r, x_f))] \\ & -\mathbb{E}_{x_f}[\log(1 - D_{Ra}(x_f, x_r))], \end{aligned} \tag{3}$$

where $D_{Ra}(\cdot)$ is the RaD formulated as $D_{Ra}(x_r, x_f) = \sigma(C(x_r) - \mathbb{E}_{x_f}[C(x_f)])$, $C(\cdot)$ means the discriminator output, $\mathbb{E}_{x_f}[\cdot]$ represents the operation of taking average for all fake data in the mini-batch, and $\sigma(\cdot)$ is the sigmoid function.

Following [15], our generator consists of three losses: the L_1, the perceptual loss L_{percep}, and the adversarial loss.

Following [27], [29], [32], [44], we use L_1 loss function to constrain the content of a generated SR image to be close to the HR image. The L_1 loss is defined in Eq. 4:

$$\begin{aligned} L_1 = & \frac{1}{Wr \times Hr \times C} \\ & \times \sum_{w=1}^{Wr} \sum_{h=1}^{Hr} \sum_{c=1}^{C} \| F_\theta^G(I_i^{LR})(w,h,c) - I_i^{HR}(w,h,c) \|_1, \end{aligned} \tag{4}$$

where $F_\theta^G(\cdot)$ represents the function of the generator, θ is the parameters of the generator and I_i means the i-th image. This function treats every position in the image equally.

The perceptual loss [45] aims to measure the perceptual similarity between the SR image and the corresponding HR image, which minimizes the distance between two high-level features extracted from a pre-trained network before the activation layers. Both the SR and HR images are taken as the input to the pre-trained VGG19 and the VGG19-54 layer features are extracted. The perceptual loss is defined as:

$$L_{percep} = \| F_\theta^{VGG}(G(I_i^{LR})) - F_\theta^{VGG}(I_i^{HR}) \|_1, \tag{5}$$

where $F_\theta^{VGG}(\cdot)$ is the features from the 4-th convolution layer before the 5-th maxpooling layer in the pre-trained VGG19 network and I_i is the i-th image, $G(\cdot)$ is the function of the generator.

The adversarial loss for the generator is in a symmetrical form against the discriminator:

$$\begin{aligned} L_G^{Ra} = & -\mathbb{E}_{x_r}[\log(1 - D_{Ra}(x_r, x_f))] \\ & -\mathbb{E}_{x_f}[\log(D_{Ra}(x_f, x_r))]. \end{aligned} \tag{6}$$

B. GLOBAL GUIDANCE MECHANISM

Inspired by the LSF-based top-down facilitation of recognition in the visual cortex [18], we deem that the LFB can guide the HFB to recover more detail information with the global context features of the input. As shown in Fig. 2, the output feature maps of the l-th RRDB in the LFB is used to guide the subsequent feature extraction of the h-th RRDB in the HFB. Specifically, the output feature maps are concatenated into the input feature maps of the h-th RRDB in the HFB. The injection of feedback information from the global level can be beneficial for the fine-grained reconstruction. We have demonstrated the effectiveness of the guidance in our experiments.

C. MASK NETWORK

To make the final reconstructed SR image focus on high-frequency details, we have embedded an attention mechanism in our framework. As presented in Fig. 2, we design a mask network to produce an attention mask for adaptively reconstructing the final output image, achieving a better trade-off between reconstruction accuracy and perceptual quality. Similar to the LFB, the mask network is stacked by M RRDBs after the shared module. Then we use the upsampling layer to upscale the attention feature map. The feature map is then processed by the sigmoid function to a probability matrix which enables the dual SR framework to yield superior results. Unlike other works that fuse the SR output of each branch to the final output [17], [46], we merge the feature maps in the process of SR image reconstruction which is before the final reconstruction convolution layer.

The feature maps extracted from the mask network module is defined as:

$$W_M = H_{mask}(F_{SF}), \tag{7}$$

where H_{mask} denotes mask network. F_{SF} is the output of shared module. The mask A can be formulated as:

$$A = \sigma(W_M), \tag{8}$$

We use the attention mask A to fuse the feature maps of the low-frequency branch and the high-frequency branch as follow:

$$I_y = (1 - A) \cdot f_{low}(F_{SF}) + A \cdot f_{high}(F_{SF}), \tag{9}$$

where $f_{high}(F_{SF})$ represents the reconstructed feature map from the HFB, $f_{low}(F_{SF})$ represents the reconstructed feature map from the LFB, and A denotes the attention mask indicating to what extent each pixel of the $f_{high}(F_{SF})$ contributes to the final output image. In this way, the mask network can learn the weight of each pixel of the feature map, leading to reconstruct the SR image with higher visual quality and less deformed textures. Furthermore, the mask network adaptively reconstructs the final output merged from the LFB and the HFB with the last convolution layer.

IV. EXPERIMENTS

In this section, we first describe our network training settings, then present the quantitative and visual results of the proposed network compared with state-of-the-art methods on benchmark datasets. To study the effects of the guidance and the dual branches in the proposed network, we conduct some ablation study experiments by removing these components and compare their differences, respectively.

TABLE 1. The numbers of RRDB and training loss functions for different modules in our GDSR.

Name	Number of RRDB	training loss
Shared Module	2	$-$
Low-Frequency Branch	15	MSE
Mask network	5	$-$
High-Frequency Branch	15	$VGG + GAN + L_1$
Final Output	$-$	L_1

A. TRAINING DETAILS

DIV2K dataset [47] contains 800, 100 and 100 images of 2K-resolution for training, validation, and testing, respectively. Following [15], [27], [32], we use 800 training images from the DIV2K dataset as the training set. At testing stage, we also use five standard benchmark datasets: Set5 [48], Set14 [49], BSD100 [50], Urban100 [51], and Manga109 [52]. The LR images are obtained by bicubic downsampling (BI) from the original high-resolution images. According to the work of Blau et al. [53], the perceptual quality of super-resolved images is not always improved with the increase of PSNR/SSIM values. We adopt the perceptual index (PI) [37] and root means square error (RMSE) as our quantitative measurements, where PI measures the perceptual quality of the super-resolved image and RMSE measures the reconstruction loss between the HR image and the SR image. Lower values of both PI and RMSE represent better quality.

The training loss functions of our network are as shown in Table 1. The training process is divided into two stages. First, we use MSE loss to pre-train a PSNR-oriented model with all branches. We then employ the trained PSNR-oriented model as an initialization for the network of the HFB. Second, we train the HFB in an adversarial manner. Meanwhile, we continue to use the L_1 loss to update the LFB and the mask network branch until the model converges.

At the training stage, the inputs are augmented by rotating and flipping. Our network is optimized with ADAM optimizer [54], whose hyper-parameters β_1 and β_2 are set to $\beta_1 = 0.9$ and $\beta_2 = 0.999$. We randomly crop HR images to 128×128 patches. Following [21], [27], [29], [32], [55], [56], the initial learning rate is set to 1×10^{-4} and then decreases to half every 2×10^5 iterations. The number of RRDBs for different modules in GDSR is set as shown in Table 1. We set the block number in the global guidance mechanism as $l = 10$ and $h = 5$. We implement our model with the PyTorch framework [57] on two NVIDIA GeForce RTX 2080Ti GPUs.

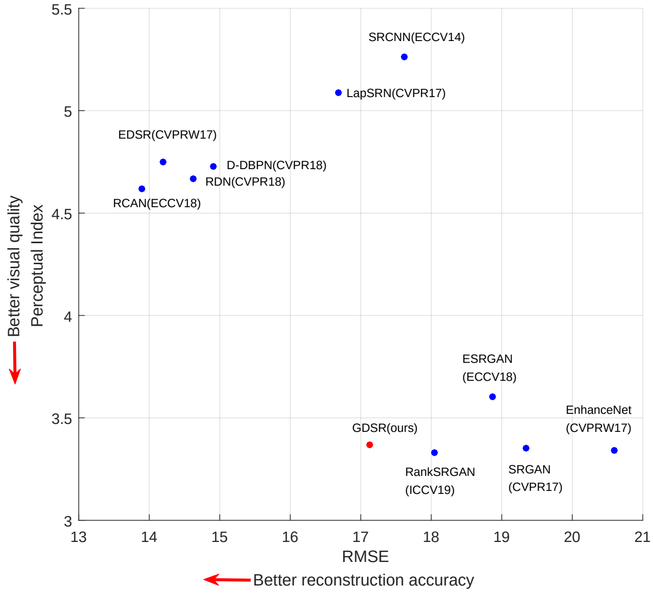


FIGURE 4. The trade-off of RMSE and PI on the Urban100 benchmark dataset of our method and state-of-the-art methods for 4× super-resolution. Among all the methods, our GDSR is the closest to the origin of the coordinates, achieving a good balance between the accuracy and perceptual quality of the SR images.

B. QUANTITATIVE COMPARISONS

As shown in Fig. 4, the methods on the top-left region are MSE-based, which have lower RMSE loss but higher PI value. They have high reconstruction accuracy but poor visual quality with over-smoothed edges. On the contrary, the methods on the bottom-right region are GAN-based, including SRGAN [33], EnhanceNet [34], ESRGAN [15], RankSRGAN [40] and our method, which have lower PI value. Although the previous GAN-based methods obtain more photo-realistic images than the MSE-based methods, they have higher RMSE loss, resulting in more deformed textures in the SR images. Our GDSR attains the lowest RMSE loss and comparatively lower PI value among all the GAN-based methods, and it can produce SR images of better perceptual quality and relatively higher reconstruction accuracy.

To further show the performance of our method more clearly, we compare the results of DNI [38] with our GDSR on the measurement index LPIPS [58]. LPIPS calculates the perceptual similarity of the images, which is recently a common measurement index to evaluate image quality in the Super-Resolution field. The evaluation result of LPIPS is more close to human perception, which provides a good trade-off between perception and distortion. We choose the MSE-based method SRResNet [33] and the GAN-based method RankSRGAN [40] to interpolate with different interpolation parameter α, which is set to 0.2, 0.4, 0.6, 0.8. As shown in Table. 3, the LPIPS of our method GDSR in

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2020.2995175, IEEE Access

TABLE 2. Quantitative results for 4× super-resolution. The best results are highlighted.

	Set5			Set14			BSD100			Urban100			Manga109		
	PSNR	SSIM	PI	PSNR	SSIM	PI	PSNR	SSIM	PI	PSNR	SSIM	PI	PSNR	SSIM	PI
ESRGAN	30.47	0.8518	**3.32**	26.29	0.6984	2.74	25.32	0.6505	2.39	24.36	0.7330	3.61	28.44	0.8599	3.19
GDSR(ours)	**30.93**	**0.8641**	3.48	**27.56**	**0.7723**	**2.72**	**26.02**	**0.6782**	**2.29**	**25.11**	**0.7557**	**3.38**	**28.95**	**0.8734**	**3.09**

TABLE 3. The quantitative comparisons of DNI between SRResNet and RankSRGAN, and GDSR model on the Urban100 benchmark dataset and Manga109 benchmark dataset. The best results are highlighted.

Method	Urban100		Manga109	
	PSNR	LPIPS	PSNR	LPIPS
SRResNet	**26.15**	0.224	**30.50**	0.108
DNI_02	25.33	0.207	28.00	0.099
DNI_04	24.70	0.187	26.69	0.090
DNI_06	24.62	0.163	26.62	0.074
DNI_08	24.76	0.144	27.85	0.065
RankSRGAN	24.49	0.139	27.89	0.075
GDSR	25.11	**0.126**	28.95	**0.061**

the two benchmark datasets Urban100 and Manga109 are lowest, which means that the image quality of our method is much better than that of simple interpolation between the PSNR-oriented model and the GAN-based model. It can be seen that the PSNR of our method is higher than that of RankSRGAN and other DNI models in the benchmark dataset Manga109, which means that our method has less distortion. In fact, our mask network can choose optimal mask parameters automatically to achieve good perception-distortion trade-off.

Furthermore, more quantitative comparison results of the performance of the proposed method with the perceptual SR methods ESRGAN [15] are listed in Table 2. The evaluation metrics include PSNR, SSIM, and PI [37]. Table 2 shows their performance on five test datasets: Set5, Set14, BSD100, Urban100, Manga109. Note that lower PI value indicates better visual quality, while higher PSNR/SSIM values mean higher reconstruction accuracy. When comparing our method with ESRGAN, we find that GDSR achieves the best PI performance on most datasets except Set5. Furthermore, the improvement of perceptual scores comes at the price of PSNR. Note that in all test sets, GDSR obtains the highest PSNR/SSIM values comparing with ESRGAN. Our proposed method achieves a lower PI value and higher PSNR/SSIM values, which means it has the better visual quality and higher reconstruction accuracy.

C. QUALITATIVE RESULTS

We compare our GDSR on some public benchmark datasets with the MSE-based methods: SRResNet [33], EDSR [27], D-DBPN [59], and the GAN-based approaches: SRGAN [33], EnhanceNet [34], ESRGAN [15], RankSRGAN [40].

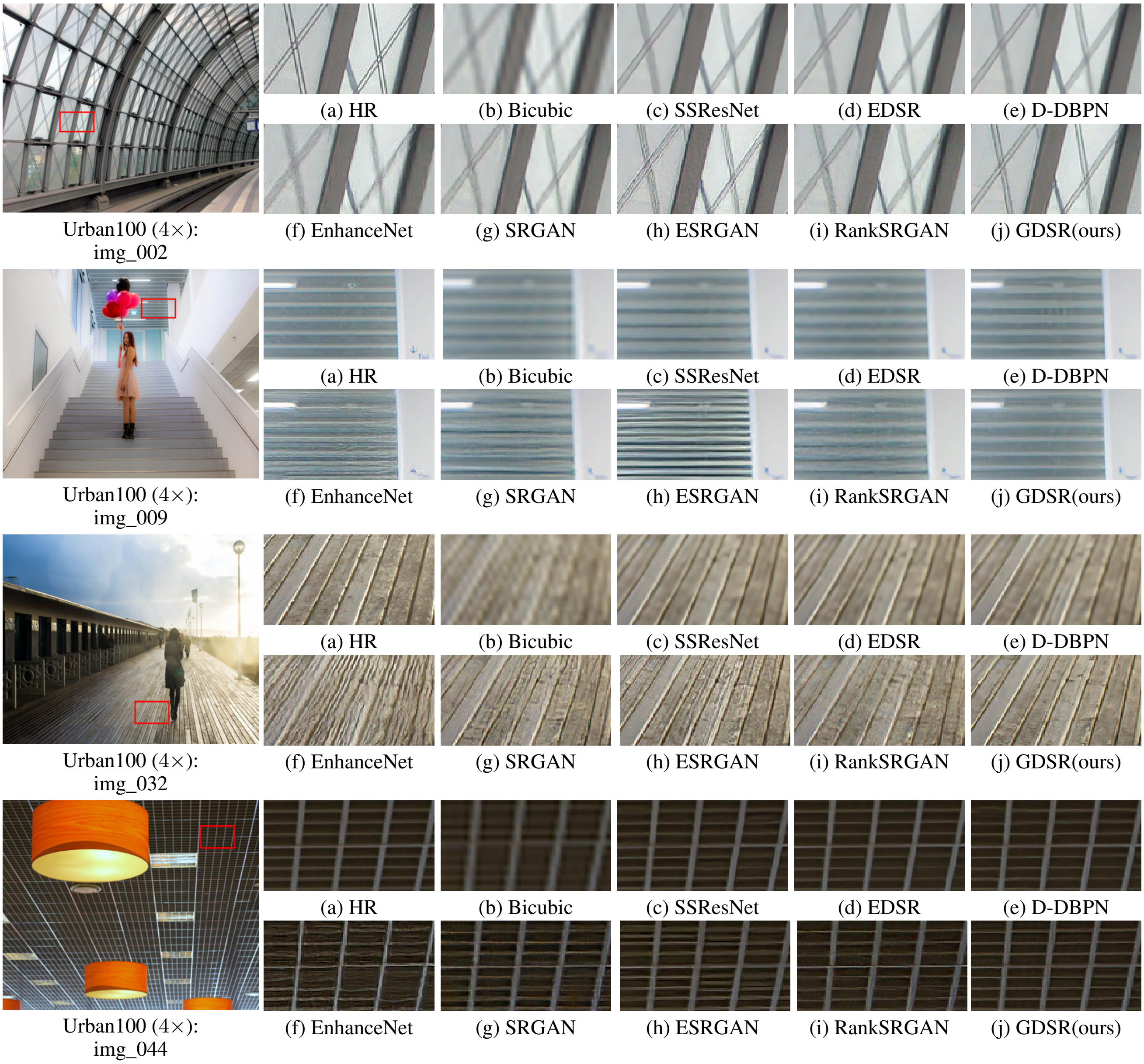
We show some visual examples, where we observe that our method could generate more realistic textures without introducing additional artifacts. As shown in Fig. 5, our proposed network outperforms other methods by a large margin in visual quality. Our network can generate images with more fine-grained textures and high-frequency details without deformation. For example, for the image 'img_002' of Urban100, the edges from MSE-based methods are over-smoothed. EnhanceNet and ESRGAN generate the streaks with noise. The margins of cropped parts of SRGAN and RankSRGAN are blurry. Our GDSR can produce clear and natural stripes of the window frame without noise. The cropped parts of the image 'img_044' in Urban100 are full of stripes. All the compared MSE-based methods suffer from blurry artifacts, failing to recover the structure and the gap of the stripes. The result of EnhanceNet is full of noisy and ESRGAN generates noise and wrong textures. The streaks of SRGAN, RankSRGAN are indistinct and blurred, while our GDSR can recover them correctly, producing more pleasing results and being faithful to the HR image. These representative comparisons demonstrate the strong ability of our GDSR for producing more photo-realistic and high-quality SR images.

We also found that our method can generate more detail lines without artifacts in bright-color background. As shown in Fig. 5, we can find that the Window frames in 'img_013' of Urban100 of SRGAN and ESRGAN are difficult to distinguish. EnhanceNet and RankSRGAN generate 'img_013' of Urban100 with too many noise. Our method GDSR can recover the lines of window frames without artifacts, which overcomes the above disadvantages of GAN-based methods.

Furthermore, we show the visual example of the DNI [38] and GDSR model in Fig. 7. We can see that the picture of DNI becomes more and more photo-realistic with the increase of α value, but those photo-realistic images are accompanied by unsatisfying artifacts and noise. The image generated by our model is more photo-realistic with less noise.

D. RESULTS WITH REAL-WORLD DATASET

We further compare our model with some others on real-world images to test the robustness of our model. We use the Nikon dataset from RealSR [60] to test, which is a testing dataset commonly used in real-world Super-Resolution filed. Several evaluation metrics exist for real-world images, such as SSEQ [61], LPIPS [58], and DIBR-Synthesized Image Quality Metric [62]. We choose SSEQ and LPIPS as our evaluation metrics where the open-source codes found online are utilized. SSEQ calculates the spatial-spectral entropy of image blocks to obtain the relationship between the image pixels. LPIPS calculates the perceptual similarity of the images. Lower values of both SSEQ and LPIPS represent better quality. From Table. 4, we can find that our GDSR performs



Urban100 (4×): img_002

(a) HR (b) Bicubic (c) SSResNet (d) EDSR (e) D-DBPN

(f) EnhanceNet (g) SRGAN (h) ESRGAN (i) RankSRGAN (j) GDSR(ours)

Urban100 (4×): img_009

(a) HR (b) Bicubic (c) SSResNet (d) EDSR (e) D-DBPN

(f) EnhanceNet (g) SRGAN (h) ESRGAN (i) RankSRGAN (j) GDSR(ours)

Urban100 (4×): img_032

(a) HR (b) Bicubic (c) SSResNet (d) EDSR (e) D-DBPN

(f) EnhanceNet (g) SRGAN (h) ESRGAN (i) RankSRGAN (j) GDSR(ours)

Urban100 (4×): img_044

(a) HR (b) Bicubic (c) SSResNet (d) EDSR (e) D-DBPN

(f) EnhanceNet (g) SRGAN (h) ESRGAN (i) RankSRGAN (j) GDSR(ours)

FIGURE 5. Visual comparison for 4× SR with BI model on Urban100 dataset.

best in compared models.

TABLE 4. The quantitative comparisons of SRResNet, ESRGAN, and GDSR model in RealSR Nikon test dataset. The best results are highlighted.

Method	Nikon	
	SSEQ	LPIPS
SRResNet	54.600	0.448
ESRGAN	45.989	0.417
GDSR	**44.824**	**0.412**

E. ABLATION STUDY

To study the effects of the structure in the proposed method, we conduct ablation experiments by removing the components and testing the differences. We remove the global guidance to verify its influences. Then we train a single branch in an adversarial manner without global guidance and mask network to verify the effect of our proposed dual branches. We also compare the differences of the final reconstruction outputs of the HFB and the dual branches in the same network. The visual comparisons are illustrated in Fig. 8 and Fig. 9. Detailed discussions are provided below.

1) Removing the Guidance

We first remove the top-down guidance in our network. An obvious performance decrease can be observed in Fig. 8. For image ‘KarappoHighschool’ in Manga109, the model without guidance introduces some unnatural noise and blurry edges, while GDSR can generate clear SR image. The HFB

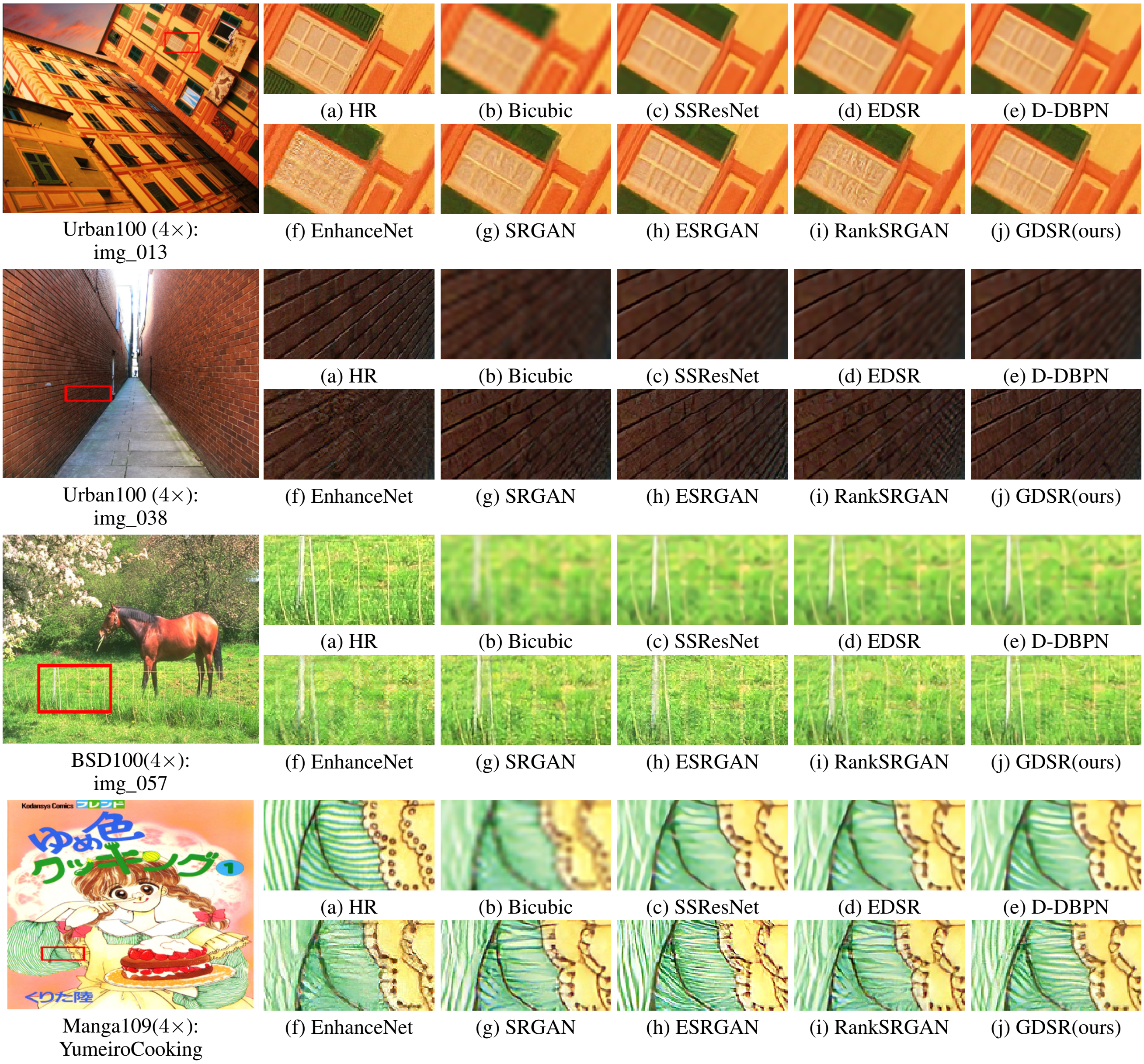


FIGURE 6. Visual comparison for 4× SR with BI model on Urban100, BSD100, and Manga109 datasets.

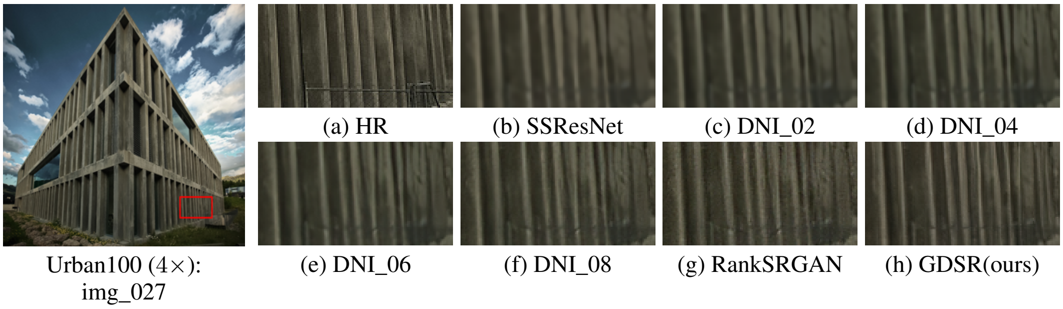
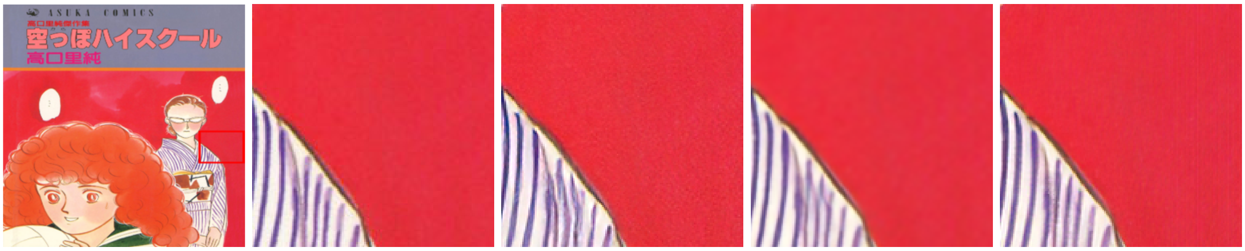
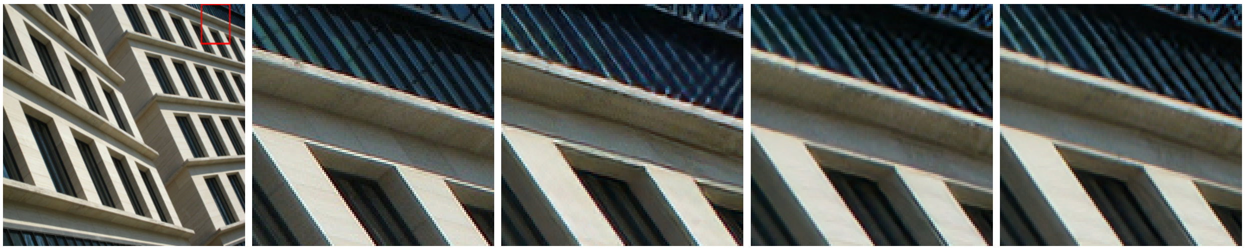


FIGURE 7. Visual comparison for 4× SR between DNI and GDSR model on Urban100 dataset.



Manga109 (4×) KarappoHighschool (a) HR (b) w/o guide HFB (c) w/o guide GDSR (d) GDSR (ours)

FIGURE 8. The visual results of the ablation study of GDSR with the guide. Without the guidance mechanism, the network tends to generate over-smoothed textures, presenting bad visual quality.



Urban100 (4×): img_025 (a) HR (b) ESRGAN_17 (c) HFB (d) GDSR (ours)

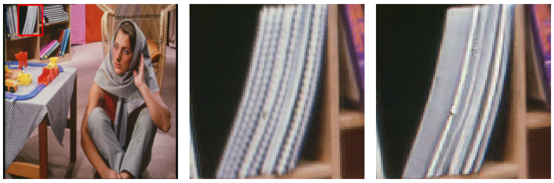
FIGURE 9. The visual results of ESRGAN_17, HFB, and GDSR. Our GDSR recovers more correct structures of the cropped part.

without the global guidance from the LFB introduces blurring artifacts. The characters in the cropped image generated by GDSR are clearer and more recognizable due to the benefit of top-down guidance. We also perform the experiment that replaces concatenation by addition of the output feature maps from the LFB to the HFB but we observe no differences.

2) Compared with Single Branch

Firstly, we train the single branch in the adversarial manner, which can be treated as ESRGAN with 17 RRDBs. Then, We generate the SR images from the HFB in the original network by adding the final reconstruction convolution layer. The experimental visual comparison results are shown in Fig. 9. We can see that GDSR outperforms the single branch ESRGAN_17 by a large margin. The single branch ESRGAN_17 tends to introduce unpleasant and unnatural artifacts. The HFB with the global guidance from the LFB can alleviate the artifacts but the lines of eaves in img_025 have additional textures. By employing the mask network to adaptively reconstruct the final output from the LFB and the HFB, our GDSR can alleviate heavy artifacts and noise to generate more correct and clearer stripes. The visual analysis indicates that our dual branches structure plays an important role in our GDSR to achieve a better trade-off between perceptual quality and reconstruction accuracy in SR images.

We also give the quantitative results of ESRGAN_17, HFB, and GDSR. As shown in Table 5, comparing with ESRGAN_17, the HFB gains higher PSNR and SSIM values than the ESRGAN_17. Our GDSR has achieved the highest PSNR and SSIM, demonstrating that our GDSR benefits from the two-branch design and reconstructs more accurate SR images.



Set14: barbara (a)GT (b)GDSR

FIGURE 10. The 4× SR visual result of 'barbara' in Set14.

TABLE 5. The quantitative comparisons of ESRGAN_17, HFB, and GDSR model in test datasets. The best results are highlighted.

Method	Set5		Set14	
	PSNR	SSIM	PSNR	SSIM
ESRGAN_17	30.17	0.8490	27.11	0.7573
HFB	30.38	0.8489	27.04	0.7568
GDSR	**30.93**	**0.8641**	**27.56**	**0.7723**

V. LIMITATIONS AND FUTURE WORK

As SISR is a serious ill-posed problem, it is unavoidable that our method has some limitations. One interesting failure on an image in the Set14 dataset is shown in Fig. 10, where the model blurs the complicated stripes visible in the HR image as smooth areas. The reason for the results is that the model does not have enough features to learn, which usually occurs between pairs of LR and HR images. The complicated stripes gradually disappear in the LR image as the size of the LR image reduces.

The model is already competitive in terms of visual results. Future work will focus on reducing the depth of the network and applying shrinking methods to speed up the model. We also want to add a temporal consistency term to use the model for video super-resolution.

VI. CONCLUSION

In this paper, we proposed a novel left-right asymmetric network for image SR to achieve a better trade-off between reconstruction accuracy and perceptual quality. We used two different training strategies to train the low-frequency branch (LFB) and the high-frequency branch (HFB), aligning with the goal to make complementary branches. The LFB is trained with MSE loss to pursue accuracy and the HFB is trained with the GAN adversarial loss to extract high-frequency features. Furthermore, we proposed a top-down guidance mechanism to guide the high-frequency feature extraction in the HFB. The high-level feature from the LFB helps the HFB to extract more high-frequency texture information. To take full advantage of both high-frequency and low-frequency features, we used a mask network to adaptively reconstruct the final output image. Our GDSR can reconstruct accurate and realistic super-resolution images, benefiting from the complementary branches to extract the high-frequency features and the low-frequency features, the guidance mechanism to guide the high-frequency feature extraction, and the mask network to fuse the features from two branches. Extensive benchmark evaluations demonstrated the effectiveness of our proposed network, which achieved superiority over state-of-the-art methods.

REFERENCES

[1] H. Seibel, S. Goldenstein, and A. Rocha, "Eyes on the target: Super-resolution and license-plate recognition in low-quality surveillance videos," IEEE access, vol. 5, pp. 20 020–20 035, 2017.

[2] K. Jiang, Z. Wang, P. Yi, G. Wang, K. Gu, and J. Jiang, "Atmfn: Adaptive-threshold-based multi-model fusion network for compressed face hallucination," IEEE Transactions on Multimedia, 2019.

[3] S. Zhang, G. Liang, S. Pan, and L. Zheng, "A fast medical image super resolution method based on deep learning network," IEEE Access, vol. 7, pp. 12 319–12 327, 2018.

[4] X. Bing, W. Zhang, L. Zheng, and Y. Zhang, "Medical image super resolution using improved generative adversarial networks," IEEE Access, vol. 7, pp. 145 030–145 038, 2019.

[5] K. Jiang, Z. Wang, P. Yi, J. Jiang, J. Xiao, and Y. Yao, "Deep distillation recursive network for remote sensing imagery super-resolution," Remote Sensing, vol. 10, no. 11, p. 1700, 2018.

[6] X. Yang, W. Wu, K. Liu, P. W. Kim, A. K. Sangaiah, and G. Jeon, "Long-distance object recognition with image super resolution: A comparative study," IEEE Access, vol. 6, pp. 13 429–13 438, 2018.

[7] X. Zhao, W. Li, Y. Zhang, and Z. Feng, "Residual super-resolution single shot network for low-resolution object detection," IEEE Access, vol. 6, pp. 47 780–47 793, 2018.

[8] Z. Wang, P. Yi, K. Jiang, J. Jiang, Z. Han, T. Lu, and J. Ma, "Multi-memory convolutional neural network for video super-resolution," IEEE Transactions on Image Processing, vol. 28, no. 5, pp. 2530–2544, 2018.

[9] P. Yi, Z. Wang, K. Jiang, Z. Shao, and J. Ma, "Multi-temporal ultra dense memory network for video super-resolution," IEEE Transactions on Circuits and Systems for Video Technology, 2019.

[10] Z. Wang, J. Chen, and S. C. Hoi, "Deep learning for image super-resolution: A survey," arXiv preprint arXiv:1902.06068, 2019.

[11] S. Anwar, S. Khan, and N. Barnes, "A deep journey into super-resolution: A survey," arXiv preprint arXiv:1904.07523, 2019.

[12] Z. Wang, A. C. Bovik, H. R. Sheikh, E. P. Simoncelli et al., "Image quality assessment: from error visibility to structural similarity," IEEE transactions on image processing, vol. 13, no. 4, pp. 600–612, 2004.

[13] Y. Zhang, K. Li, K. Li, L. Wang, B. Zhong, and Y. Fu, "Image super-resolution using very deep residual channel attention networks," in Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 286–301.

[14] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in Advances in neural information processing systems, 2014, pp. 2672–2680.

[15] X. Wang, K. Yu, S. Wu, J. Gu, Y. Liu, C. Dong, Y. Qiao, and C. C. Loy, "ESRGAN: Enhanced super-resolution generative adversarial networks," in European Conference on Computer Vision. Springer, 2018, pp. 63–79.

[16] J. Pan, S. Liu, D. Sun, J. Zhang, Y. Liu, J. Ren, Z. Li, J. Tang, H. Lu, Y.-W. Tai et al., "Learning dual convolutional neural networks for low-level vision," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 3070–3079.

[17] J. Qin, Z. Xie, Y. Shi, and W. Wen, "Difficulty-aware image super resolution via deep adaptive dual-network," arXiv preprint arXiv:1904.05802, 2019.

[18] C. Cheng, Y. Fu, Y.-G. Jiang, W. Liu, W. Lu, J. Feng, and X. Xue, "Dual skipping networks," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 4071–4079.

[19] C. Dong, C. C. Loy, K. He, and X. Tang, "Image super-resolution using deep convolutional networks," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 38, no. 2, pp. 295–307, 2016.

[20] C. Dong, C. C. Loy, and X. Tang, "Accelerating the super-resolution convolutional neural network," in European conference on computer vision. Springer, 2016, pp. 391–407.

[21] W. Shi, J. Caballero, F. Huszár, 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 Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 1874–1883.

[22] M. D. Zeiler, D. Krishnan, G. W. Taylor, and R. Fergus, "Deconvolutional networks." in Proceedings of the IEEE conference on computer vision and pattern recognition, vol. 10, 2010, p. 7.

[23] M. D. Zeiler, G. W. Taylor, R. Fergus et al., "Adaptive deconvolutional networks for mid and high level feature learning." in ICCV, vol. 1, no. 2, 2011, p. 6.

[24] J.-H. Kim and J.-S. Lee, "Deep residual network with enhanced upscaling module for super-resolution," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2018, pp. 800–808.

[25] J. Kim, J. K. Lee, and K. M. Lee, "Accurate image super-resolution using very deep convolutional networks," in 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 2016, pp. 1646–1654.

[26] J. Kim, J. Kwon Lee, and K. Mu Lee, "Deeply-recursive convolutional network for image super-resolution," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 1637–1645.

[27] B. Lim, S. Son, H. Kim, S. Nah, and K. Mu Lee, "Enhanced deep residual networks for single image super-resolution," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2017, pp. 136–144.

[28] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 770–778.

[29] W.-S. Lai, J.-B. Huang, N. Ahuja, and M.-H. Yang, "Deep laplacian pyramid networks for fast and accurate super-resolution," in IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 5835–5843.

[30] A. Shocher, N. Cohen, and M. Irani, "'Zero-shot' super-resolution using deep internal learning," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 3118–3126.

[31] K. Zhang, W. Zuo, and L. Zhang, "Learning a single convolutional super-resolution network for multiple degradations," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 3262–3271.

[32] Y. Zhang, Y. Tian, Y. Kong, B. Zhong, and Y. Fu, "Residual dense network for image super-resolution," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 2472–2481.

[33] C. Ledig, L. Theis, F. Huszár, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, Z. Wang et al., "Photo-realistic single image super-resolution using a generative adversarial network," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 4681–4690.

[34] M. S. Sajjadi, B. Scholkopf, and M. Hirsch, "EnhanceNet: Single image super-resolution through automated texture synthesis," in Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 4491–4500.

[35] L. A. Gatys, A. S. Ecker, and M. Bethge, "Image style transfer using convolutional neural networks," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 2414–2423.

[36] A. Jolicoeur-Martineau, "The relativistic discriminator: a key element missing from standard GAN," arXiv preprint arXiv:1807.00734, 2018.

[37] Y. Blau, R. Mechrez, R. Timofte, T. Michaeli, and L. Zelnik-Manor, "The 2018 PIRM challenge on perceptual image super-resolution," in European Conference on Computer Vision. Springer, 2018, pp. 334–355.

[38] X. Wang, K. Yu, C. Dong, X. Tang, and C. C. Loy, "Deep network interpolation for continuous imagery effect transition," in The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 2019.

[39] K. Jiang, Z. Wang, P. Yi, G. Wang, T. Lu, and J. Jiang, "Edge-enhanced gan for remote sensing image superresolution," IEEE Transactions on Geoscience and Remote Sensing, vol. 57, no. 8, pp. 5799–5812, 2019.

[40] W. Zhang, Y. Liu, C. Dong, and Y. Qiao, "RankSRGAN: Generative adversarial networks with ranker for image super-resolution," arXiv preprint arXiv:1908.06382, 2019.

[41] H. Kuang, H. Wang, X. Ma, and X. Liu, "Image super-resolution based on dual path network," in 2018 10th International Conference on Measuring Technology and Mechatronics Automation (ICMTMA). IEEE, 2018, pp. 225–228.

[42] Y. Chen, J. Li, H. Xiao, X. Jin, S. Yan, and J. Feng, "Dual path networks," in Advances in neural information processing systems, 2017, pp. 4467–4475.

[43] W. Han, S. Chang, D. Liu, M. Yu, M. Witbrock, and T. S. Huang, "Image super-resolution via dual-state recurrent networks," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 1654–1663.

[44] W. Lai, J. Huang, N. Ahuja, and M. Yang, "Fast and accurate image super-resolution with deep Laplacian pyramid networks," IEEE Transactions on Pattern Analysis and Machine Intelligence, pp. 1–14, Aug 2018.

[45] J. Johnson, A. Alahi, and L. Fei-Fei, "Perceptual losses for real-time style transfer and super-resolution," in European conference on computer vision. Springer, 2016, pp. 694–711.

[46] A. Pumarola, A. Agudo, A. M. Martinez, A. Sanfeliu, and F. Moreno-Noguer, "Ganimation: Anatomically-aware facial animation from a single image," in Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 818–833.

[47] R. Timofte, E. Agustsson, L. Van Gool, M.-H. Yang, and L. Zhang, "NTIRE 2017 challenge on single image super-resolution: Methods and results," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2017, pp. 114–125.

[48] M. Bevilacqua, A. Roumy, C. Guillemot, and M.-L. Alberi Morel, "Low-complexity single-image super-resolution based on nonnegative neighbor embedding," in Proceedings of the British Machine Vision Conference. BMVA Press, 2012, pp. 135.1–135.10.

[49] R. Zeyde, M. Elad, and M. Protter, "On single image scale-up using sparse-representations," in Proceedings of the 7th International Conference on Curves and Surfaces. Berlin, Heidelberg: Springer-Verlag, 2012, pp. 711–730.

[50] D. Martin, C. Fowlkes, D. Tal, J. Malik et al., "A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics," in ICCV, 2001.

[51] J.-B. Huang, A. Singh, and N. Ahuja, "Single image super-resolution from transformed self-exemplars," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 5197–5206.

[52] Y. Matsui, K. Ito, Y. Aramaki, A. Fujimoto, T. Ogawa, T. Yamasaki, and K. Aizawa, "Sketch-based manga retrieval using manga109 dataset," Multimedia Tools and Applications, vol. 76, no. 20, pp. 21 811–21 838, 2017.

[53] Y. Blau and T. Michaeli, "The perception-distortion tradeoff," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 6228–6237.

[54] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in ICLR, 2015.

[55] C. Liu, X. Sun, C. Chen, P. L. Rosin, Y. Yan, L. Jin, and X. Peng, "Multi-scale residual hierarchical dense networks for single image super-resolution," IEEE Access, vol. 7, pp. 60 572–60 583, 2019.

[56] Y. Zhang, K. Li, K. Li, B. Zhong, and Y. Fu, "Residual non-local attention networks for image restoration," in ICLR, 2019.

[57] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer, "Automatic differentiation in pytorch," in Proc. Adv. Neural Inf. Process. Syst. Workshop Autodiff, Dec., 2017, pp. 1–4.

[58] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in CVPR, 2018.

[59] M. Haris, G. Shakhnarovich, and N. Ukita, "Deep back-projection networks for super-resolution," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 1664–1673.

[60] J. Cai, H. Zeng, H. Yong, Z. Cao, and L. Zhang, "Toward real-world single image super-resolution: A new benchmark and a new model," in Proceedings of the IEEE International Conference on Computer Vision, 2019.

[61] L. Liu, B. Liu, H. Huang, and A. C. Bovik, "No-reference image quality assessment based on spatial and spectral entropies," Signal Processing: Image Communication, vol. 29, no. 8, pp. 856–863, 2014.

[62] G. Wang, Z. Wang, K. Gu, L. Li, Z. Xia, and L. Wu, "Blind quality metric of dibr-synthesized images in the discrete wavelet transform domain," IEEE Transactions on Image Processing, vol. 29, pp. 1802–1814, 2019.