APSIPA Transactions on Signal and Information Processing, 2022, 11, e29

Original Paper

ITS-Net: Iterative Two-Stream Network for Image Super-Resolution

Wei Li[1,2], Yan Huang[1]*, Yilong Yin[1] and Jingliang Peng[3,4]

[1] *School of Software, Shandong University, Jinan, China*
[2] *IOT Research Center, China Electronics Standardization Institute, Beijing, China*
[3] *Shandong Provincial Key Laboratory of Network Based Intelligent Computing, University of Jinan, Jinan, China*
[4] *School of Information Science and Engineering, University of Jinan, Jinan, China*

ABSTRACT

Remarkable progress on single image super-resolution (SISR) has been achieved with deep convolutional neural network (CNN) based approaches. These methods usually divide the images into high-frequency (HF) and low-frequency (LF) components and mainly recover the high-frequency component in a supervised manner. However, only simple interpolation manners, such as bilinear and bicubic, are utilized in the LF components recovery process, which limits the performance of these SISR approaches. We argue that the recovery of LF components also plays important roles in SISR, and to relieve the problem, we propose an iterative two-stream network (ITS-Net) which recovers the LF and HF components with convolution operations, respectively, thus better high-resolution images can be obtained. To achieve this, we utilize a sub-network with convolution and deconvolution operations to recover the LF components, and an iterative learning strategy is used to obtain well recovered LF and HF components. Extensive experiments on

*Corresponding author: Yan Huang, yan.h@sdu.edu.cn.

Received 13 April 2022; Revised 22 May 2022
ISSN 2048-7703; DOI 10.1561/116.00000018

various benchmarking datasets demonstrate the effectiveness of our approach comparing with state-of-the-art CNN based approaches.

Keywords: Image resolution, low-frequency, high-frequency, deep learning, iterative learning.

1 Introduction

Researchers in many areas have taken single image super-resolution (SISR) as fundamental technology, for it plays important roles in many applications, such as virtual reality [2, 5], augmented reality [17, 26] and 3D reconstruction [6, 23]. SISR is a fundamental low-level computer vision problem, which aims to reconstruct a desired high-resolution (HR) image with pleasing visual quality from its low-resolution (LR) counterpart. However, it is a typical ill-posed problem due to the fact that lots of detailed information would be lost during the down-sampling process. Hence, how to recover a high-quality HR image is still a challenging task.

Impressive progress has been achieved by deep convolutional neural networks (DCNN) [7, 18] in SISR, which divide the HR images into HF and LF components, and mainly focus on recovering the high frequency components of the HR images. In these approaches, such as Li *et al.* [15], the LF components are usually recovered by simple interpolation manners, such as bilinear or bicubic, and the final HR images are obtained by adding the HF and LF components with convolutional operations. We would argue that the recoveries of LF components also play important roles in SISR and the estimation of LF components with simple interpolation manners which is insufficient for current state-of-the-art approaches. Figure 1 demonstrates the general pipeline of residual learning framework of current SISR approaches. The final HR images can be obtained by adding the recovered HF components and LF components

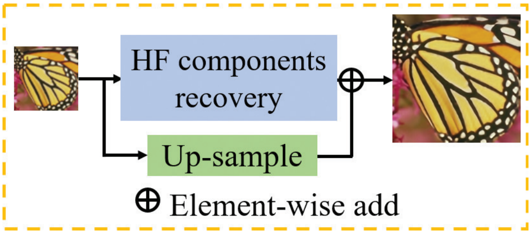

Figure 1: The framework of residual learning strategy. Up-sample means commonly used interpolations operations, such as bilinear or bicubic.

(the up-sampled LR input), and interpolation operations, such as bicubic or bilinear, are commonly used as the up-sampling operations for the LR input. However, the LF information is not fully exploited, and it is insufficient to describe LF components recovery with interpolation operations, which limits the performance of these approaches.

Based on the above considerations, we propose an iterative two-stream network (ITS-Net) to recover both LF and HF components for SISR, which uses a simple but effective sub-net to replace the interpolation-based manners (bicubic or bilinear) for LF components recovery. For the HF components branch, state-of-the-art approaches have proven their effectiveness in SISR, hence, we directly use the network structures that are proposed in Chao *et al.* [4], Kim *et al.* [11], Li *et al.* [15] and Mei *et al.* [22] to validate the effectiveness of our strategy. Besides, to effectively combine the LF and HF recovery branches, we propose an iterative two-stream network to update the LF and HF branches separately, thus the reconstruction ability of the network can be intrinsically improved. In specific, instead of interpolation operation (bicubic or bilinear), we propose to use convolutional operations for LF components recovery, thus more effective information can be recovered and better HR images can be obtained. Besides, an iterative learning strategy is used to sufficiently recover the LF and HF components, which guarantees that the proposed framework can be convergent. As shown in Figure 2, our ITS-Net approach shows better visual quality in comparison with other state-of-the-art SISR methods.

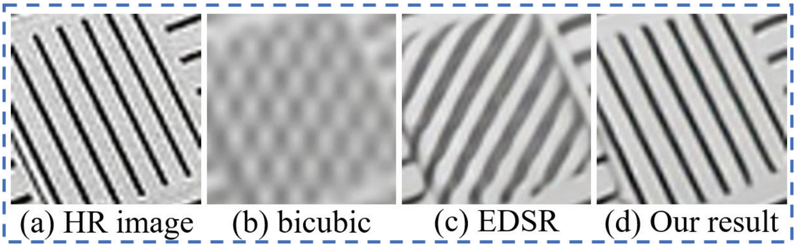

Figure 2: The comparisons with state-of-the-art approaches: (a) ground truth HR image; (b) bicubic; (c) EDSR [16]; (d) our result.

The contributions of our work for SISR can be summarized as follows:

- We propose a novel iterative two-stream network (ITS-Net) for SISR, which contains two sub-networks to recover both low-frequency (LF) and high-frequency (HF) components separately, thus better HR images can be recovered. The sub-network of LF component branch only contains several convolutional layers which is simple and effective.

- We propose an iterative learning strategy to sufficiently train the two-stream network, and the LF and HF sub-networks can refine each other to learn more contextual information, thus the reconstruction performance is intrinsically improved.

- Extensive experiments demonstrate the superiority of the proposed ITS-Net compared with other state-of-the-art SISR methods.

2 Related Work

Recently, deep learning based image SR technology has been rapidly developed in image SR. Chao *et al.* [4] introduced CNN into the SR task and proposed the SRCNN model with a three-layer network to learn the mapping from LR images to HR images. This model achieved better performance compared with the traditional algorithms. Then, Kim *et al.* [11] proposed the VDSR model that used a very deep network with 20 layers to improve the SR performance. Lai *et al.* [14] proposed the lapSRN method that took the original LR images as input and progressively reconstructed the sub-band residuals of HR images. To reuse the features of each layer in CNN, Huang *et al.* [9] proposed a dense convolutional network (DenseNet) by connecting each layer to every other layer in a feed-forward fashion. Song *et al.* [27] proposed a channel attention based iterative residual learning for depth map super-resolution. Liu *et al.* [18] proposed a residual feature aggregation (RFA) framework for effective feature extraction. Tai *et al.* [28] proposed a deep recursive residual Network (DRRN) with a very deep CNN model (up to 52 convolutional layers) for SISR.

Deeper neural networks usually generate much more parameters which occupy a huge amount of storage resources and suffer from overfitting. Recurrent structure can effectively reduce the parameters of the network and gain better generalization power for image SR. Li *et al.* [15] proposed a SRFBN framework by exploring the feasibility of feedback mechanism for image SR. Although it designed a RNN with feedback connections to deliver the highest-level features to a shallow layer, it fails to fully use high-level features captured under large receptive fields. To learn more useful information, Guo *et al.* [7] proposed a dual regression network for single image super-resolution by learning an additional dual regression mapping to estimate the down-sampling kernel and reconstruct LR images with effective supervision. Yang *et al.* [31] proposed a texture transformer network for image super-resolution. Ma *et al.* [19] proposed restoring high-resolution gradient maps by a gradient branch to provide additional structure priors for the SR process and imposed a second-order restriction with gradient loss. Mei *et al.* [22] proposed a Non-Local Sparse Attention (NLSA) with dynamic sparse attention pattern for image super-resolution. Pesavento *et al.* [25] proposed an attention-based multi-reference super-resolution network. Meanwhile, to sufficiently extract and fuse different levels of features, Kong *et al.* [13], Wang *et al.* [29], and Xie *et al.* [30] propose effective frameworks, where features with different levels of detail, including high-frequency, medium frequency and low frequency, are processed discriminatively to efficiently recover HR images.

The above literature review reveals that a significant improvement on SISR has been achieved by deep learning based methods. However, we find that existing methods just focus on HF information and ignore the significance of LF information for SISR.

3 Our Approach

3.1 Overview

Residual learning strategies have proven the effectiveness in single image super-resolution (SISR) [11, 15], and these approaches mainly focus on recovering the HF component, while the LF component is neglected. In this paper, we argue that the recovery of LF component also works positively in SISR. Therefore, we propose a novel ITS-Net, which can recover the HF and LF respectively. In specific, taking the LR image I_l as input, our ITS-Net (M_{ITS}) contains two sub-networks for image SR, the LF sub-network for LF components (I^{LF}) recovery and HF sub-network for HF components (I^{HF}) recovery. Then, the LF and HF components are combined for the estimation of HR image (I_h).

Figure 3(a) shows the pipeline of our ITS-Net, and the main difference with state-of-the-art residual learning approaches [15] is that we use a sub-network to replace the commonly used bilinear or bicubic interpolations for LF components recovery.

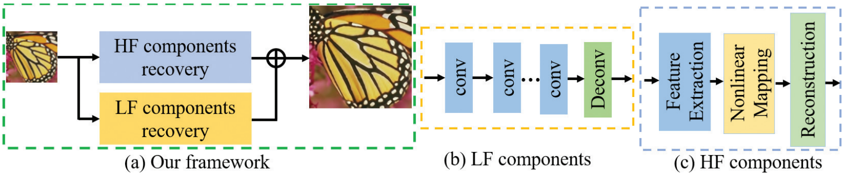

Figure 3: (a) Shows the pipeline of the proposed ITS-Net, which contains two sub-networks for HF and LF components recovery. (b) and (c) Illustrate the sub-networks for HF and LF recovery, respectively.

The process can be formulated as:

$$\begin{aligned} I^{LF} &= M_{LF}(I_l) \\ I^{HF} &= M_{HF}(I_l) \\ I_h &= I^{HF} + I^{LF} \end{aligned} \tag{1}$$

where M_{LF} and M_{HF} are LF and HF components recovery operations, respectively. I_l and I_h mean the input LR and estimated HR images.

3.2 LF Sub-network

In this work, for the branch of LF components estimation M_{LF}, we use convolutional operations instead of commonly used interpolation methods (bilinear or bicubic) to obtain more effective LF information.

Figure 3(b) illustrates the structure of our LF sub-network M_{LF}, which includes two operations. Inspired by Chao *et al.* [4] and Kim *et al.* [11], the proposed M_{LF} contains N convolution layers (named C, kernel size of $s \times s$ and m channels) and a deconvolution layer (named DC). We use $Conv(s, m_i, m_o)$ and $Deconv(s, m_i, m_o)$ to denote a convolution and a deconvolution layer respectively, where s is the size of the filter, m_i is the number of input channels and m_o is the number of the output channels. The sub-network of LF can be formulated as:

$$\begin{aligned} f_1^{LF} &= C_1(I_l) \\ f_t^{LF} &= C(f_1^{LF}) \\ I^{LF} &= DC(f_t^{LF}) \end{aligned} \tag{2}$$

where C_1 means one convolutional operation with $Conv(s, 3, m_o)$, C means (N-2) convolutional operations with $Conv(s, m_i, m_o)$, and DC means the deconvolutional operation with $Deconv(s, m_i, m_o)$, and pixelshuffle operation can also be used here. f_1^{LF} and f_t^{LF} are the features obtained by C_1 and C operations. I^{LF} means the obtained results of the LF component recovery branch.

We will provide more analysis of the number of layers N of the branch of LF components estimation in the experiment.

3.3 HF Sub-network

For the HF component recovery, we directly use the well-verified network structures which are utilized in state-of-the-art approaches. To sufficiently evaluate the effectiveness of our proposed structure, the HF recovery structures utilized in SRCNN [4], VDSR [11], SRFBN [15], and NLSN [22] are evaluated here. As demonstrated in Figure 3, we combine the LF component recovery network with the HF recovery structures.

Figure 3(c) illustrates the pipeline of HF component recovery. Following SRCNN [4], to recover high-resolution images, three modules are utilized, including feature extraction module (F_{FE}), nonlinearing mapping module (F_{NP}) and image reconstruction module (F_{IR}). Inspired by SRCNN [4], VDSR [11], SRFBN [15], and NLSN [22] also contain F_{FE}, F_{NP} and F_{IR} for HF components recovery, and the mainly differences among these approaches are F_{NP}. In specific, VDSR utilizes very deep network structure in F_{NP}. SRFBN exploits feedback network to extract effective HF information for high performance. NLSN utilizes a Non-Local Sparse Attention operation and ResBlocks in the HF recovery process. Please see SRCNN [4], VDSR [11], SRFBN [15] and NLSN [22] for more details of the corresponding F_{NP}. In this paper, for simplicity, the recovery of HF component can be formulated as:

$$\begin{aligned} f_{FE} &= F_{FE}(I_l) \\ f_{NP} &= F_{NP}(f_{FE}) \\ I^{HF} &= F_{IR}(f_{NP}) \end{aligned} \tag{3}$$

where f_{FE} and f_{NP} are obtained features of F_{FE} and F_{NP}, and I^{HF} is the recovery HF component. And the final recovery HR image can be obtained by $I_h = I^{LF} + I^{HF}$.

3.4 Iterative Learning Strategy and Loss

In this section, we provide more details of the proposed iterative learning strategy and the loss function used in our paper. Note that $L1$ loss is used in this paper.

Our ITS-Net (M_{ITS}) contains LF component recovery branch (M_{LF}) and HF component recovery branch (M_{HF}), and the training process contains three stages:

(1) With fixed M_{HF}, the M_{LF} is first trained to obtain pre-trained model of M_{LF}, and the loss function is defined as:

$$l_{LF} = ||I_{GT} - I^{LF}||_1 \tag{4}$$

where I_{GT} is the HR ground truth and I^{LF} is the output of M_{LF}.

(2) The M_{ITS} is trained with pre-trained M_{LF} obtained in (1), and the parameters of sub-network M_{LF} is fixed, which aims to optimize M_{HF}, and the loss function is defined as:

$$\begin{aligned} I_h &= I^{LF} + I^{HF} \\ l &= ||I_{GT} - I_h||_1 \end{aligned} \tag{5}$$

where I^{LF} and I^{HF} are the the output of M_{LF} and M_{HF}, respectively. I_h means the final recovered HR image.

(3) The M_{ITS} is finetuned without fixing M_{LF} and M_{HF}, and the parameters of sub-network M_{LF} and M_{HF} are initialized with the parameters obtained in (2), the loss function is same with Eq. 5.

Note that the iterative learning strategy can be regarded as a residual learning process. In the first step, only LF recovery branch is utilized, which aims to recover coarse HR images, and as proved by previous approaches, the recovered coarse HR images are better than results obtained by interpolation operations. To make it better to understand, we define the results obtained by LF branch as LF components. In the second step, the residual between coarse HR images and ground truth can be obtained, and in such process, information of edges and structures can be mainly refined which can be regarded as learning HF components.

3.5 Implementation Details

To effectively train our ITS-Net, we initialize the network parameters with the strategies in He *et al.* [8], kernel size with 3 × 3 is utilized in our paper, and Kingma and Ba [12] is employed to optimize the parameters of the network. As described in Section 3.4, the ITS-Net contains three stages. In the first stage, we set the learning rate as 1×10^{-4}, and 200 epochs are used in this stage. In the second stage, we set the learning rate as 1×10^{-4}, and the learning rate multiplies by 0.5 for every 200 epochs, and 600 epochs are employed. While for the third stage, we set learning rate as 1.25×10^{-5}, and 200 epochs are employed.

4 Experiment

To prove the effectiveness of our approach, quantitative and qualitative comparisons are provided between our ITS-Net with state-of-the-art approaches on different datasets.

4.1 Experimental Settings

Datasets: Following [24], DIV2K [1] is used as our training data, and to make full use of training data. Five standard benchmark datasets: Set5 [3], Set14 [32], BSDS100 [20], Urban100 [10] and Manga109 [21] are used to validate the performance of our ITS-Net approach.

We use bicubic down-sampling (BI) as the standard degradation model to generate LR images from ground truth HR images. Besides, to verify the effectiveness of our ITS-Net, following [15, 16], we also use BD as a degradation model which first applies Gaussian blur to HR images, then down-samples the HR images to LR images with bicubic degradation. In our experiments, we use 7 × 7 sized Gaussian kernel with standard deviation 1.6 for blurring.

Evaluation metrics: Two commonly used image quality metrics PSNR and SSIM are used as the metrics. PSNR is commonly used to quantify reconstruction quality for images and SSIM can be used for measuring the structure similarity between two images.

Baseline approaches: We demonstrate the effectiveness of ITS-Net by comparing it with the following state-of-the-art image super-resolution methods, including: Bicubic interpolation, SRCNN [4], VDSR [11], DRRN [28], IRCANN [33], EDSR [16], SRFBN [15], HAN [24], and NLSN [22]. Note that HAN [24] uses parameters trained by RCAN [34] to initialize the network, and better performance can be obtained, and in this paper, to compare equally, we re-trained HAN [24] without using parameters of RCAN [34] to initialize the network.

4.2 Quantitative Results

In this section, we provide quantitative comparison results between our approach and state-of-the-art approaches on various datasets under up-sampling scales of ×3 and ×4 with BI and BD degradation, respectively. Tables 1 and 2 illustrate the quantitative comparison results of different datasets with BI and BD degradations and under up-sampling scales of ×4 and ×3, respectively. We can see that remarkable performance can be achieved by current SISR approaches, such as EDSR, SRFBN, HAN and NLSN. Note that different strategies are utilized in approaches of HAN [24] and RCAN [34], where attention and residual learning strategies are first used in the LR features before up-sampling the LR features, and the LR component recovery strategy used in

Table 1: Quantitative comparison results between our approach and state-of-the-art approaches under up-sampling scale of ×4 for BI degradation. The best is shown in bold, and the second is shown in underline. The improvements are shown in blue. The higher the better.

Methods ×4	Set5	Set14	BSDS100	Urban100	Manga109
	PSNR/SSIM	PSNR/SSIM	PSNR/SSIM	PSNR/SSIM	PSNR/SSIM
Bicubic	28.42/0.8104	26.00/0.7027	25.96/0.6675	23.14/0.6577	24.89/0.7866
DRRN [28]	31.68/0.8888	28.21/0.7721	27.38/0.7284	25.44/0.7638	29.18/0.8914
EDSR [16]	32.46/0.8968	28.80/0.7876	27.71/0.7420	26.64/0.8033	31.02/0.9148
HAN [24]	32.54/0.8995	28.83/0.7875	27.72/0.7412	26.74/0.8065	31.17/0.9164
SRCNN [4]	30.48/0.8628	27.50/0.7513	26.90/0.7101	24.52/0.7221	27.58/0.8555
SRCNN[4] + our	30.91/0.8748	27.92/0.7614	27.15/0.7174	25.05/0.7349	28.19/0.8699
improvements	0.43/0.0120	0.42/0.0101	0.25/0.0073	0.53/0.0128	0.61/0.0144
VDSR [11]	31.35/0.8830	28.02/0.7680	27.29/0.7260	25.18/0.7540	28.83/0.8870
VDSR[11] + our	31.54/0.8894	28.20/0.7740	27.41/0.7292	25.31/0.7586	29.01/0.8932
improvements	0.19/0.0064	0.18/0.0060	0.12/0.0032	0.13/0.0046	0.18/0.0062
SRFBN [15]	32.47/0.8983	28.81/0.7868	27.72/0.7409	26.60/0.8015	31.15/0.9160
SRFBN[15] + our	32.55/0.8999	28.87/0.7884	27.80/0.7421	26.73/0.8031	31.23/0.9178
improvements	0.09/0.0016	0.06/0.0016	0.08/0.0012	0.07/0.0016	0.08/0.0018
NLSN [22]	32.59/0.9000	28.87/0.7891	27.78/0.7444	26.96/0.8109	31.27/0.9184
NLSN[22] + our	**32.63/0.9016**	**28.92/0.7908**	**27.84/0.7456**	**27.01/0.8119**	**31.31/0.9196**
improvements	0.04/0.0016	0.05/0.0017	0.06/0.0012	0.05/0.0010	0.04/0.0012

Table 2: Quantitative comparison results between our approach and state-of-the-art approaches under up-sampling scale of ×3 for BD degradation. The best is shown in bold, and the second is shown in underline. The improvements are shown in blue. The higher the better.

Methods ×4	Set5	Set14	BSDS100	Urban100	Manga109
	PSNR/SSIM	PSNR/SSIM	PSNR/SSIM	PSNR/SSIM	PSNR/SSIM
Bicubic	28.34/0.8161	26.12/0.7106	26.02/0.6733	23.20/0.6661	25.03/0.7987
IRCANN [33]	33.38/0.9182	29.63/0.8281	28.65/0.7922	26.77/0.8154	31.15/0.9245
HAN [24]	34.67/0.9287	30.51/0.8455	29.22/0.8074	28.52/0.8587	34.10/0.9471
SRCNN [4]	31.63/0.8888	28.52/0.7924	27.76/0.7526	25.31/0.7612	28.79/0.8851
SRCNN [4] + our	31.95/0.9043	28.87/0.8096	28.13/0.7721	25.59/0.7798	29.17/0.9045
improvements	0.32/0.0155	0.35/0.0172	0.37/0.0195	0.28/0.0186	0.38/0.0194
VDSR [11]	33.30/0.9159	29.67/0.8269	28.63/0.7903	26.75/0.8145	31.66/0.9260
VDSR [11] + our	33.42/0.9197	29.82/0.8299	28.76/0.7964	26.86/0.8187	31.81/0.9298
improvements	0.12/0.0038	0.15/0.0030	0.13/0.0031	0.11/0.0042	0.15/0.0038
SRFBN [15]	34.66/0.9283	30.48/0.8439	29.21/0.8069	28.48/0.8581	34.07/0.9466
SRFBN [15] + our	34.71/0.9293	30.55/0.8452	29.32/0.8081	28.57/0.8594	34.15/0.9480
improvements	0.05/0.0010	0.07/0.0013	0.11/0.0012	0.09/0.0013	0.08/0.0014
NLSN [22]	34.72/0.9295	30.57/0.8455	29.34/0.8083	28.58/0.8595	34.16/0.9482
NLSN[22] + our	**34.77**/**0.9304**	**30.61**/**0.8463**	**29.40**/**0.8091**	**28.63**/**0.8602**	**34.20**/**0.9491**
improvements	0.05/0.0009	0.04/0.0008	0.06/0.0008	0.05/0.0007	0.04/0.0009

this paper can not be combined in these approaches. Therefore, in this paper, we mainly evaluate our approaches with structures utilized in SRCNN [4], VDSR [11], SRFBN [15] and NLSN [22].

As shown in Tables 1 and 2, the results with our learning strategy are with better PSNR and SSIM, which proves the effectiveness of our approach. Specifically, SRCNN [4] and VDSR [11] can be largely improved by our learning strategy with degradations of BI and BD under up-sampling scales of ×3 and ×4. As shown in Table 1, the improvement of SRCNN on Urban100 dataset [10] is 0.53 in PSNR evaluation metric with BI degradation under up-sampling scale of ×4, while the corresponding improvement on Urban100 dataset [10] is 0.28 in PSNR evaluation metric with BD degradation under up-sampling scale of ×3 (as shown in Table 2). For VDSR [11], the average improvements with our learning strategy are 0.16 and 0.132 in PSNR with BI and BD degradations, respectively.

Meanwhile, for HF recovery frameworks with better performances, such as SRFBN [15] and NLSN [22], as demonstrated in Tables 1 and 2, stable improvements in PSNR and SSIM among all of the evaluation datasets can also be obtained with our proposed strategies (with SRFBN [15] and NLSN [22] as HF recovery frameworks), which proves that the sub-network of LF components works positively to recover better HR images for different degradation models. Note that the parameters of model used in HAN [24] is about 16 M, while the count of parameters of SRFBN [15] + our is only about 4 M. Hence, our ITS-Net shows the advantages with higher performance and fewer model parameters. Besides, the stable improvements of our ITS-Net with different HF recovery frameworks can also proves the robust compatible propriety of our approach with different degradation models.

4.3 Qualitative Results

Figure 4 illustrates the qualitative comparison results between our approach and state-of-the-art approaches under up-sampling scale of ×4 on Manga109 dataset [21] and Urban100 [10] dataset with BI degradation. We can find that results obtained by bicubic interpolation suffers heavily blurring, and approaches of (c) to (g) in Figure 4 can recover better HR images. Meanwhile, comparing with (b) to (f), more visual appealing results with sharper boundaries and more details can be obtained by our approach, which also demonstrates that the sub-network of LF components can improve the performance of HR images with BI degradation.

Meanwhile, Figure 5 shows the qualitative comparison results between our approach and state-of-the-art approaches under up-sampling scale of ×3 on Urban100 [10] dataset with BD degradation, where the HR images are first applied with Gaussian blur, then the input LR images can be obtained by down-sampling the HR images with bicubic degradation. We can clearly see

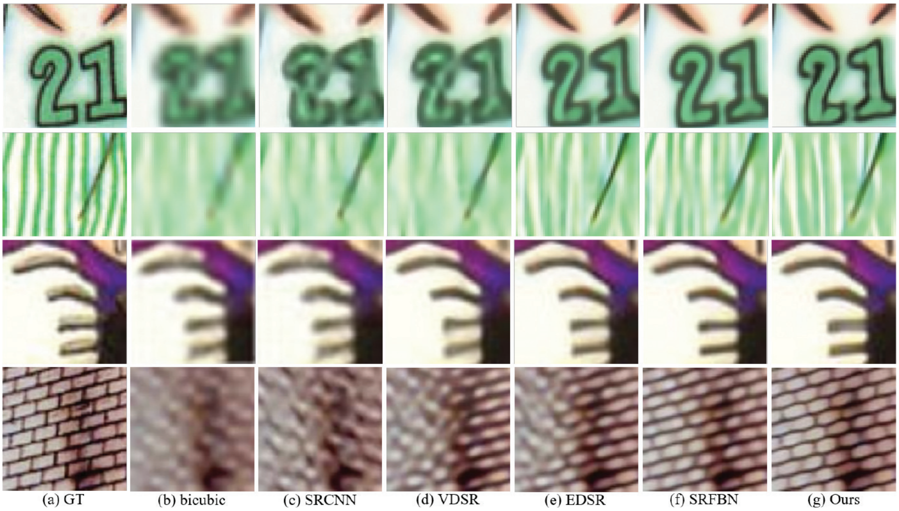

Figure 4: Qualitative comparison results with state-of-the-art approaches under up-sampling scale of ×4 with BI degradation on Manga109 dataset [21] (the first row) and Urban100 [10] dataset (the second row). (a) Ground truth HR image; (b) bicubic; (c) SRCNN [4]; (d) VDSR [11]; (e) EDSR [16]; (f) SRFBN [15]; and (g) SRFBN + our.

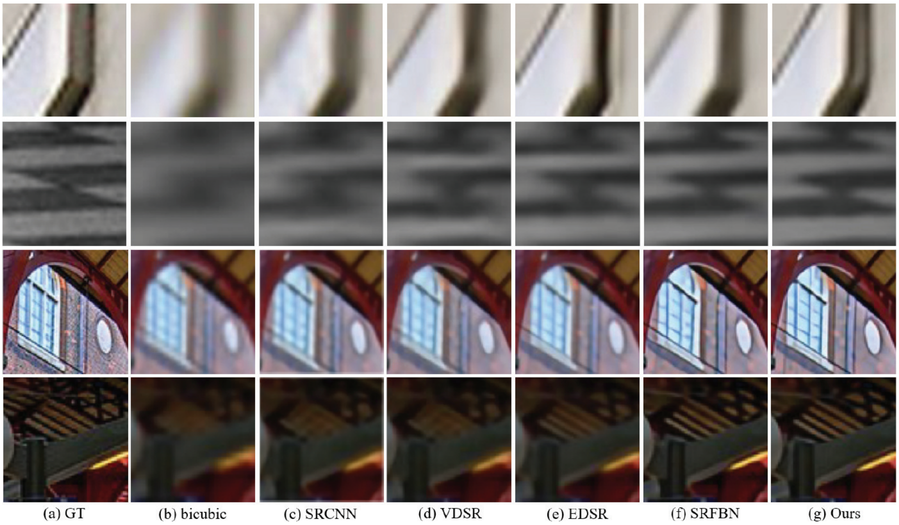

Figure 5: Qualitative comparison results with state-of-the-art approaches under up-sampling scale of ×3 with BD degradation on Urban100 [10] dataset. (a) Ground truth HR image; (b) bicubic; (c) SRCNN [4]; (d) VDSR [11]; (e) EDSR [16]; (f) SRFBN [15]; and (g) SRFBN + our.

that more details can be recovered by ITS-Net, and more visually appealing results can also be obtained by our approach under BD degradation model.

4.4 Ablation Study

In this section, we provide more analysis of the different number of layers in sub-network of LF components (with VDSR [11]) on Set5 dataset [3] under up-sampling scale of ×4 with BI degradation, and the PSRN results are demonstrated in Table 3. To reduce the training time, we use VDSR [11] as the baseline here. We can see that as the number of layers increases, better PSNR can be obtained, and the performance will be convergent. Hence, we set the number of layers in sub-network of LF components as 5 in our paper. As discussed in our paper, in the LF component recovery branch, only 5 convolution layers are utilized, the parameter size of the LF branch is only about 0.1 M, which has little influence on the memory and evaluation time.

Besides, to show the effectiveness of the proposed iterative learning strategy, we also provide ablation study here. As shown in Table 4, in comparison with VDSR [11], directly training VDSR [11] with LF component branch can improve the performance (VDSR + LF (joint)) (with 0.11 db improvement in PSNR). Meanwhile, better PSNR results can be obtained by our iterative learning strategy (VDSR + LF (iter)) with 0.13 db improvement in PSNR,

Table 3: Results of $PSNR$ with different number of layers in sub-network of low-frequency on Set5 dataset [3].

Number	1	3	5	7
PSNR	34.48	34.49	34.54	34.54

Table 4: Results of $PSNR$ with training strategies on Set5 dataset [3]. VDSR + LR (joint) means result obtained by training VDSR wtih LR component recovery branch directly, and VDSR + LR (iter) means result obtained by training VDSR wtih LR component recovery branch with our iterative training strategy.

Methods	VDSR	VDSR + LF (joint)	VDSR + LF (iter)
PSNR	31.30	31.41	31.54

which proves the effectiveness of the proposed iterative learning strategy. Moreover, the proposed ITS-Net scheme with LF sub-network and iterative learning strategy to learn LF and HF components is flexible and can be easily combined with other effective frameworks to improve performance for SISR.

5 Conclusion

In this paper, we propose an iterative two-stream network (ITS-Net) for accurate image SR, which consists of HF and LF components recovery branches. Note that the LF recovery branch is simple but effective which is utilized to replace the commonly used up-sampling operations (bicubic or bilinear), thus the representation of LF components can be successfully enriched and better HR images can be recovered. Meanwhile, to improve the compatible ability, the iterative training strategy is employed which achieves more accurate image SR performance by iteratively refining the high-frequency sub-network and the low-frequency sub-network. Extensive experiments demonstrated that table improvements can be obtained by our approach with different HF recovery strategies, which proves the superiority of our proposed ITS-Net.

References

[1] E. Agustsson and R. Timofte, “Ntire 2017 Challenge on Single Image Super-resolution: Dataset and Study”, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2017, 126–35.

[2] C. Anthes, R. J. García-Hernández, M. Wiedemann, and D. Kranzlmüller, "State of the Art of Virtual Reality Technology", in *2016 IEEE Aerospace Conference*, IEEE, 2016, 1–19.

[3] M. Bevilacqua, A. Roumy, C. Guillemot, and M. L. Alberi-Morel, "Low-complexity Single-image Super-resolution based on Nonnegative Neighbor Embedding", 2012.

[4] D. Chao, C. L. Chen, K. He, and X. Tang, "Learning a Deep Convolutional Network for Image Super-Resolution", in *ECCV*, 2014, 184–99.

[5] P. Fuchs, G. Moreau, and P. Guitton, *Virtual Reality: Concepts and Technologies*, CRC Press, 2019.

[6] A. Geiger, J. Ziegler, and C. Stiller, "Stereoscan: Dense 3D Reconstruction in Real-time", in *2011 IEEE Intelligent Vehicles Symposium (IV)*, Ieee, 2011, 963–8.

[7] Y. Guo, J. Chen, J. Wang, Q. Chen, J. Cao, Z. Deng, Y. Xu, and M. Tan, "Closed-loop Matters: Dual Regression Networks for Single Image Super-Resolution", in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[8] K. He, X. Zhang, S. Ren, and J. Sun, "Delving Deep into Rectifiers: Surpassing Human-level Performance on Imagenet Classification", in *IEEE International Conference on Computer Vision*, 2015, 1026–34.

[9] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely Connected Convolutional Networks", in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, 4700–8.

[10] J.-B. Huang, A. Singh, and N. Ahuja, "Single Image Super-resolution from Transformed Self-exemplars", in *IEEE Conference on Computer Vision and Pattern Recognition*, 2015, 5197–206.

[11] J. Kim, J. Kwon Lee, and K. Mu Lee, "Accurate Image Super-resolution Using Very Deep Convolutional Networks", in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, 1646–54.

[12] D. P. Kingma and J. Ba, "Adam: A Method for Stochastic Optimization", *arXiv preprint arXiv:1412.6980*, 2014.

[13] X. Kong, H. Zhao, Y. Qiao, and C. Dong, "Classsr: A General Framework to Accelerate Super-resolution Networks by Data Characteristic", in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, 12016–25.

[14] W. Lai, J. Huang, N. Ahuja, and M. Yang, "Deep Laplacian Pyramid Networks for Fast and Accurate Super-Resolution", *Computer Vision and Pattern Recognition*, 2017, 5835–43.

[15] Z. Li, J. Yang, Z. Liu, X. Yang, G. Jeon, and W. Wu, "Feedback Network for Image Super-resolution", in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, 3867–76.

[16] 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, 136–44.

[17] H. Ling, "Augmented Reality in Reality", *IEEE MultiMedia*, 24(3), 2017, 10–15.

[18] J. Liu, W. Zhang, Y. Tang, J. Tang, and G. Wu, "Residual Feature Aggregation Network for Image Super-Resolution", in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[19] C. Ma, Y. Rao, Y. Cheng, C. Chen, J. Lu, and J. Zhou, "Structure-Preserving Super Resolution With Gradient Guidance", in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, 7766–75, DOI: 10.1109/CVPR42600.2020.00779.

[20] 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 *IEEE International Conference on Computer Vision.* Vol. 2, IEEE, 2001, 416–23.

[21] 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*, 76(20), 2017, 21811–38.

[22] Y. Mei, Y. Fan, and Y. Zhou, "Image Super-Resolution with Non-local Sparse Attention", in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, 3517–26.

[23] E. Mouragnon, M. Lhuillier, M. Dhome, F. Dekeyser, and P. Sayd, "Real Time Localization and 3D Reconstruction", in *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, Vol. 1, IEEE, 2006, 363–70.

[24] B. Niu, W. Wen, W. Ren, X. Zhang, L. Yang, S. Wang, K. Zhang, X. Cao, and H. Shen, "Single Image Super-resolution via a Holistic Attention Network", in *European Conference on Computer Vision*, Springer, 2020, 191–207.

[25] M. Pesavento, M. Volino, and A. Hilton, "Attention-based Multi-Reference Learning for Image Super-Resolution", in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, 14697–706.

[26] D. Schmalstieg and T. Hollerer, *Augmented Reality: Principles and Practice*, Addison-Wesley Professional, 2016.

[27] X. Song, Y. Dai, D. Zhou, L. Liu, W. Li, H. Li, and R. Yang, "Channel Attention Based Iterative Residual Learning for Depth Map Super-Resolution", in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[28] Y. Tai, J. Yang, and X. Liu, “Image Super-resolution via Deep Recursive Residual Network”, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, 3147–55.

[29] L. Wang, X. Dong, Y. Wang, X. Ying, Z. Lin, W. An, and Y. Guo, “Exploring Sparsity in Image Super-resolution for Efficient Inference”, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, 4917–26.

[30] W. Xie, D. Song, C. Xu, C. Xu, H. Zhang, and Y. Wang, “Learning Frequency-aware Dynamic Network for Efficient Super-resolution”, in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, 4308–17.

[31] F. Yang, H. Yang, J. Fu, H. Lu, and B. Guo, “Learning Texture Transformer Network for Image Super-Resolution”, in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[32] R. Zeyde, M. Elad, and M. Protter, “On Single Image Scale-up Using Sparse-representations”, in *International Conference on Curves and Surfaces*, Springer, 2010, 711–30.

[33] K. Zhang, W. Zuo, S. Gu, and L. Zhang, “Learning Deep CNN Denoiser Prior for Image Restoration”, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, 3929–38.

[34] 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, 286–301.