Hierarchical Generative Adversarial Networks for Single Image Super-Resolution

Weimin Chen[1]*, Yuqing Ma[2]*, Xianglong Liu[2]†, Yi Yuan[1]
[1]NetEase Fuxi AI Lab, Hangzhou, China
[2]State Key Lab of Software Development Environment, Beihang University, China
{chenweimin, yuanyi}@corp.netease.com, {mayuqing, xlliu}@nlsde.buaa.edu.cn

Abstract

Recently, deep convolutional neural network (CNN) have achieved promising performance for single image super-resolution (SISR). However, they usually extract features on a single scale and lack sufficient supervision information, leading to undesired artifacts and unpleasant noise in super-resolution (SR) images. To address this problem, we first propose a hierarchical feature extraction module (HFEM) to extract the features in multiple scales, which helps concentrate on both local textures and global semantics. Then, a hierarchical guided reconstruction module (HGRM) is introduced to reconstruct more natural structural textures in SR images via intermediate supervisions in a progressive manner. Finally, we integrate HFEM and HGRM in a simple yet efficient end-to-end framework named hierarchical generative adversarial networks (HSRGAN) to recover consistent details, and thus obtain the semantically reasonable and visually realistic results. Extensive experiments on five common datasets demonstrate that our method shows favorable visual quality and superior quantitative performance compared to state-of-the-art methods for SISR.

1. Introduction

Single image super-resolution (SISR), which has received great attention through past few years, aims to reconstruct the high-resolution (HR) image from its low-resolution (LR) counterpart. SISR is an ill-posed problem, because there may be multiple HR images matching the same LR image through various degradation processes. A number of SISR methods have been proposed in the literature, such as interpolation-based methods [35], reconstruction-based methods [34], traditional learning-based methods [31, 32, 25, 26], and recent deep learning-based methods [3, 4, 13, 27, 11, 27, 28, 36, 21].

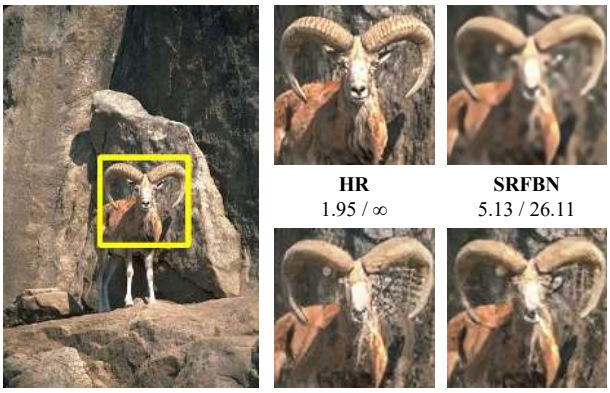

Figure 1. Visual quality from different types of SISR methods. Compared with our HSRGAN, SRFBN [12] reconstructs a over-smoothing image while SRGAN [11] introduces some undesirable noise at right horn. Our HSRGAN can reduce unpleasant artifacts and produce more convincing textures.

The early SISR methods are based on interpolation kernels [35], such as bilinear or bicubic, which are simple and efficient for real-time application. However, they can hardly effectively restore the high frequency information, and easily leads to blurred images. Furthermore, the reconstruction methods [34] introduced the prior knowledge of the image as a constraint, formulating the SISR problem as an inverse process modeling of the imaging system, such as deblurring, upper sampling and denoising. In the past decades, a number of traditional learning-based methods [31, 32, 25, 26] have been proposed to predict missing high-frequency details in LR images, by learning the mapping relationship between LR and HR images, through sparse coding [31, 32], neighbor embedding [25, 26], etc.

More recently, deep neural networks have shown the promising power in computer vision tasks, such as recogni-

*The first two authors contributed equally.
†Xianglong Liu is the corresponding author.

tion [2], classification [7], etc. Therefore, researchers adopt deep learning algorithms to solve the SISR problem. Depending on the training objectives of the model, deep learning methods can be broadly divided into two categories: 1) Former methods concentrate on minimizing the Mean Square Error (MSE) or Mean Absolute Error (MAE) between the super-resolution (SR) image and the ground-truth image, which may lead to blurriness and make the reconstructed image unrealistic [30]. 2) To recover more sharp texture details in SR image and make it look realistic, recent researchers introduce GAN loss [5, 18] into model training. Nevertheless, due to the fact that these works only extract features on a single scale and lack sufficient supervision information, they struggled to recover the local details and complex structures and thus largely suffered undesired artifacts in the SR images, as shown in Figure 1.

To alleviate these problems, we propose a simple yet efficient hierarchical generative adversarial networks (HSRGAN) for SISR, which could recover consistent details in an hierarchical manner, and thus obtain the semantically reasonable and visually realistic SR results. Specifically, we propose a hierarchical feature extraction module (HFEM) to extract the features of multiple scales using a multi-branch architecture, which helps our network concentrate on both local textures and global semantics. Furthermore, we propose a hierarchical guided reconstruction module (HGRM), where we divide the SR task of a large upscale factor into a sequence of easier sub-tasks with small upscale factors. The upscaling sub-tasks provide more convincing supervision information to help the final upscaling procedure to obtain more realistic texture details. Through extracting multi-scale information and gradually generating SR images from small to large, our proposed HSRGAN is capable of recovering the visually realistic and semantically reasonable images from a single LR images.

In summary, our main contributions are listed as follows:

- We propose the simple yet efficient hierarchical generative adversarial networks (HSRGAN) for SISR, which can stably generate SR images by taking hierarchical features and supervision information into consideration.
- We devise a hierarchical feature extraction module (HFEM) to capture the hierarchical features in multiple scales, which helps our model concentrate on both local textures and global semantics.
- We further present a hierarchical guided reconstruction module (HGRM) to exploit rich supervision information about the structural textures and reconstruct the image in a progressive manner.
- Experiments demonstrate that our method significantly outperforms other state-of-the-art methods in terms of both quantitative metrics and visual quality.

2. Related work

Since the deep learning methods have outperformed most conventional SR methods, in this section, we mainly focus on deep learning algorithms for SISR. There are mainly two directions to optimize the SISR problem. At first, most of the methods use pixel-wise MSE (L2) loss or MAE (L1) loss as objective function to reduce the distortion error. And recently, pursuing better perceptual quality of SR images has become a new trend with the boom of generative adversarial networks [5, 18].

2.1. Distortion-oriented methods

A variety of deep learning techniques have been used in SISR problem and several different designs of models have been adopted. SRCNN [3, 4] was a pioneer work that established a three-layer convolutional neural network to directly reconstruct HR image from corresponding interpolated LR one. EDSR [13] stacked modified residual blocks which removed the batch normalization layers and ReLU activation from ResNet [6] to work with the SR task. To exploit channel-wise relationships in feature maps, inspired from SENet [8], RCAN [36] introduced the residual channel attention block to improve performance. Other recent works like MemNet [23] and RDN [37] employed dense blocks [9] and SRFBN [12] adopted feedback mechanism, which commonly exists in human visual system, to refine low-level representations with high-level information.

However, the distortion-based methods mentioned above use pixel-wise MSE (L2) loss or MAE (L1) loss as the objective function, which causes blurriness or over-smoothing in SR image and makes it look unreal.

2.2. Perception-oriented methods

Since generative adversarial networks (GANs) [5, 18] was proposed, models based on GAN loss have been used in several aspects of computer vision because of more realistic details the model can generate, which is what the previous methods optimized by L1 loss or L2 loss cannot achieve. GANs employ two components, namely a generator and a discriminator, to combat with each other. The generator tries to create an SR image that the discriminator cannot distinguish from a real HR image. In this manner, SR images with more realistic textures are generated.

SRGAN [11] firstly introduced GANs into SISR problem, where the generator was composed of residual blocks [6]. To improve the naturalness of the images, perceptual and adversarial losses were used to train the model in SRGAN. Compared with the global normalization of BN layer, SFTGAN [27] proposed a spatial feature transform (SFT) layer, which used the segmentation probability maps as a reference, had a stronger ability to deal with the borders

of different objects. In this way, the effect of texture reconstruction for different objects is improved. Some other networks, such as EnhanceNet [19] and SRFeat [17], used multiple loss terms or discriminators to improve the performance. ESRGAN [28] was a variant of SRGAN. It enhanced its performance by removing the batch normalization layers and employed relativistic discriminator [10] in training. Furthermore, NatSR [21] defined the naturalness prior in the low-level domain and constrained the output image in the natural manifold, which eventually generated more natural and realistic images.

However, with the help of GANs, it may produce some unpleasant noise or unnatural textures when the network amplifies the image resolution.

3. Proposed method

We propose a hierarchical model to solve the SISR problem, named HSRGAN, which could recover consistent details in the generated images, and thus generate the semantically reasonable and visually realistic SR results. As shown in Figure 2, our hierarchical model architecture is based on the GANs, where the generator consists of two consecutive parts: the hierarchical feature extraction module and the hierarchical guided reconstruction module. For the discriminator in our models, we employ relativistic discriminator [10] in stabilizing and accelerating training procedure, as ESRGAN [28] did. We will elaborate the details of our proposed HSRGAN in the following sections.

3.1. Framework

Given an input LR image $\mathbf{I}_{\mathrm{LR}}$, the goal of our proposed model is to train a generative model that can generate the SR image $\mathbf{I}_{\mathrm{SR}}$ close to original HR image $\mathbf{I}_{\mathrm{HR}}$.

Most of the previous SISR methods extract features by stacking convolution layers with a single kernel size, which may limit the representative capability of the network. To enhance the SR performance, we apply HFEM to exploit detail textures and global information. Specifically, the input $\mathbf{I}_{\mathrm{LR}}$ is firstly fed into HFEM to extract feature maps in individual receptive field:

$$\mathbf{F}_{\mathrm{H}} = \mathbb{E}\left(\mathbf{I}_{\mathrm{LR}}\right), \tag{1}$$

where $\mathbb{E}$ represents the HFEM, which consists of multi-branch network (MBN) and feature fusion network (FFN). MBN can extract multi-scale features by several branches and the features is then fused by FFN, which further exploits the intent information.

Furthermore, the GAN-based methods may produce undesired artifacts [15], especially in the SISR problem with large magnification, due to the lack of sufficient supervision information. To handle this issue, we proposed hierarchical guided reconstruction module (HGRM) to reconstruct the final output $\mathbf{I}_{\mathrm{SR}}$, which upsamples the mixed features in a progressive manner with multiple supervision by:

$$\mathbf{I}_{\mathrm{SR}} = \mathbb{U}\left(\mathbf{F}_{\mathrm{H}}\right), \tag{2}$$

where $\mathbb{U}$ stands for the HGRM. The detail of HFEM and HGRM will be shown in Section 3.2 and 3.3.

3.2. Hierarchical feature extraction

In human visual systems, when recognizing an object, we need to pay attention to the global information as well as the local details. In the same way, suitable feature representations are crucial for networks to understand an image. The latest SISR methods extract features and achieve impressive results. But they usually extract features of LR images by convolution layers with a single kernel size, which means that they only focus on one single scale, limiting the capacity of feature representation of the network. Thus there is still much room to treat features as a set of different components and combine both global semantics and local textures.

To address these problems, inspired by [29], we utilize the HFEM to capture different levels of features from input LR images, which considers features in various scales. Our HFEM contains two components: multi-branch network (MBN) and future fusion network (FFN). Firstly, the input LR image is fed into the MBN to parallelly extract features $\mathbf{F}_{\mathrm{M}}$ in various scales:

$$\begin{aligned}\mathbf{F}_{\mathrm{M}} &= \mathbb{E}_{\mathrm{M}}(\mathbf{I}_{\mathrm{LR}}) \\ &= \left[\mathbb{E}_{\mathrm{M}}^{(1)}(\mathbf{I}_{\mathrm{LR}}), \mathbb{E}_{\mathrm{M}}^{(2)}(\mathbf{I}_{\mathrm{LR}}), \cdots, \mathbb{E}_{\mathrm{M}}^{(B)}(\mathbf{I}_{\mathrm{LR}})\right],\end{aligned} \tag{3}$$

where $\mathbb{E}_{\mathrm{M}}$ represents the MBN of our HFEM, B denotes the number of branches and $\mathbb{E}_{\mathrm{M}}^{(b)}$ means the b-th branch, $1 \leq b \leq B$. The convolution kernel size within each single branch is the same and differs from other branches. Therefore, different branches have distinct receptive fields. The branches with smaller receptive fields focus on the local textures of LR images and the branches with larger receptive fields integrates more surrounding information such as spatial relationship.

Secondly, since the concatenated features may be complementary or mutually exclusive, and the network cannot make full use of this information, we fuse the hierarchical features by FFN to further learn the intent information. Our FFN consists of a channel aggregation layer and a series of residual blocks. We feed the concatenation of the hierarchical features $\mathbf{F}_{\mathrm{M}}$ into the FFN to jointly learn the final feature representation of the input $\mathbf{I}_{\mathrm{LR}}$:

$$\mathbf{F}_{\mathrm{H}} = \mathbb{E}_{\mathrm{F}}\left(\mathbf{F}_{\mathrm{M}}\right), \tag{4}$$

where $\mathbb{E}_{\mathrm{F}}$ represents the FFN of our HFEM.

In practice, considering the performance and efficiency, we set the number of parallel branches $B = 3$ for default

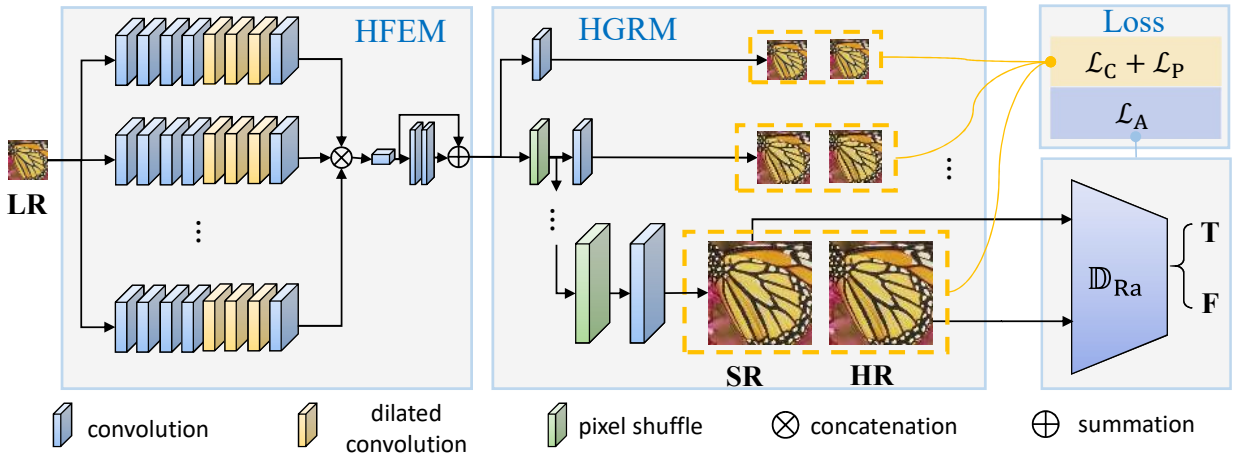

Figure 2. Architecture of our proposed HSRGAN. The bias item and activation layer are omitted for simplicity.

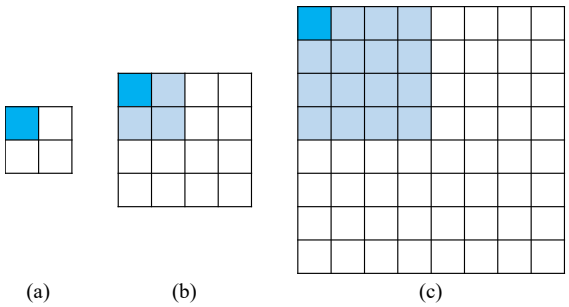

Figure 3. Illustration of pixel filling in SISR. Suppose that the pixel at left upper corner (blue) in *2-by-2* image, as shown in (a), is directly map to left upper corner in its upscale versions. (b) For a magnification of 2, we only need to fill the remaining 3 pixels (light blue). (c) For a magnification of 4, we need to fill 15 pixels (light blue), which is more likely to produce unreal pixels (image details) than for the magnification of 2.

with kernel size of 3, 5 and 7, respectively. In addition, we employ dilated convolution [33] layers to further expand receptive field while maintaining low computation cost. More details will be discussed in Section 4.3.

3.3. Hierarchical guided reconstruction

Compared with image inpainting which estimates suitable pixel information to fill holes in images, SISR is a task of filling several new sensible pixels around existing pixels, see Figure 3. As the magnification increases, the number of pixels to be filled doubles. However, most of the previous SISR methods employ a network to directly generate a SR image with a large upscale factor. Compared with the problem with a small upscale factor, the network will create more pixels based on one pixel, which may force the generator to produce a lot of unreal image details. To solve this drawback, we propose a hierarchical guided reconstruction module (HGRM) to recover the final SR image in an easy-to-hard way.

Different from conventional SISR methods, our hierarchical guided reconstruction module (HGRM) introduces more supervision information into the model. It contains T branches, where $T \geq 1$, including one main branch and $T-1$ intermediate branches. The main branch reconstructs the LR image to the target resolution and the rest branches produce the intermediate SR images with corresponding upscale factors to provide more supervision information:

$$\mathbf{I}_{\mathrm{SR}}^{(t)} = \mathbb{U}^{(t)}\left(\mathbf{F}_{\mathrm{H}}\right), \tag{5}$$

where $\mathbf{I}_{\mathrm{SR}}^{(t)}$ represents the t-th output image and $\mathbb{U}^{(t)}$ denotes the t-th branch, $1 \leq t \leq T$. It should be noted that the $\mathbb{U}^{(t)}$ shares common upscale networks (pixel shuffle layers) with its previous branches to ensure the supervision information. The main branch generates the final SR image $\mathbf{I}_{\mathrm{SR}}$ by

$$\mathbf{I}_{\mathrm{SR}} = \mathbf{I}_{\mathrm{SR}}^{(T)} = \mathbb{U}^{(T)}\left(\mathbf{F}_{\mathrm{H}}\right). \tag{6}$$

The main purpose to introduce hierarchical branches into the reconstruction module is to provide more supervision information about the image content (*e.g.* structural texture), by penalizing the network with losses between the outputs of intermediate branches and intermediate HR images generated from the ultimate HR image. Due to the fact that employing adversarial loss may force the generator to produce sharp but incorrect details, we disable the adversarial loss to intermediate branches to avoid wrong supervision information. Thus the intermediate branches mainly concentrate on the content information, while the final branch preserves the structure information and generate realistic results.

In addition, the intermediate branches can be used for combating gradient vanishing problem, and thus stabilize the generative process through providing more regulariza-

tion [22]. It is worth noting that they are used for training only and NOT used in testing or inference time.

In our experiments, we employ two branches ($T = 2$) to super-resolve the $4\times$ images, including the main branch and one intermediate branch with magnification of 2. More details will be discussed in Section 4.3 and *Supplementary Materials*.

3.4. Formulation

Given a training pair $\{\mathbf{I}_{\mathrm{LR}}, \mathbf{I}_{\mathrm{HR}}\}$, we generate $T-1$ intermediate HR images $\left\{\mathbf{I}_{\mathrm{HR}}^{(1)}, \cdots, \mathbf{I}_{\mathrm{HR}}^{(T-1)}\right\}$ from its corresponding HR image $\mathbf{I}_{\mathrm{HR}}$ by bicubic kernel. Specifically, $\mathbf{I}_{\mathrm{HR}} = \mathbf{I}_{\mathrm{HR}}^{(T)}$.

As the Figure 2 shows, we define the content loss, perceptual loss and adversarial loss to optimize the parameters of our model, which is similar to [28]. The content loss function of our model is:

$$\mathcal{L}_{\mathrm{C}} = \frac{1}{T}\sum_{i=1}^{T}\left\|\mathbf{I}_{\mathrm{SR}}^{(i)} - \mathbf{I}_{\mathrm{HR}}^{(i)}\right\|_1, \tag{7}$$

where $\|\cdot\|_1$ represents the 1-norm distance operator.

The perceptual loss can be expressed as:

$$\mathcal{L}_{\mathrm{P}} = \frac{1}{T}\sum_{i=1}^{T}\left\|\phi\left(\mathbf{I}_{\mathrm{SR}}^{(i)}\right) - \phi\left(\mathbf{I}_{\mathrm{HR}}^{(i)}\right)\right\|_1, \tag{8}$$

where $\phi(\cdot)$ indicates the feature map obtained by the VGG19 network [20].

The adversarial loss is defined based on the enhanced discriminator of Relativistic GAN [10], denoted as $\mathbb{D}_{\mathrm{Ra}}$. Due to our idea of HGRM that intermediate branches mainly provide the supervision information about the image content, we disable the adversarial loss on the branches which may generate incorrect details and only apply it to the final SR image. The adversarial loss for generator is expressed as:

$$\begin{aligned}\mathcal{L}_{\mathrm{A}} = &-\log\left(1 - \mathbb{D}_{\mathrm{Ra}}\left(\mathbf{I}_{\mathrm{HR}}, \mathbf{I}_{\mathrm{SR}}\right)\right)\\ &-\log\left(\mathbb{D}_{\mathrm{Ra}}\left(\mathbf{I}_{\mathrm{SR}}, \mathbf{I}_{\mathrm{HR}}\right)\right).\end{aligned} \tag{9}$$

To summarize, the total loss of the generator is then defined as:

$$\mathcal{L}_{\mathrm{G}} = \mathcal{L}_{\mathrm{C}} + \lambda\mathcal{L}_{\mathrm{P}} + \eta\mathcal{L}_{\mathrm{A}}, \tag{10}$$

where λ and η are the trade-off coefficients to balance different loss terms.

4. Experiment

4.1. Dataset and settings

For training, we use the mixture of DIV2K [24] and Flickr2K [24], named DF2K. DIV2K is a high-quality (2K resolution) dataset specially organized for SISR tasks. It contains 800 images for training, 100 images for validation and 100 images for testing. Flickr2K is collected on the Flickr website, which consists of 2650 2K high resolution images. Furthermore, we enrich our training set with OutdoorSceneTraining (OST) [27], which contains rich natural textures in 7 categories. We randomly crop 128×128 patches to feed into our models with random horizontal flips and 90 degree rotations as ESRGAN [28] did.

To comprehensively evaluate our proposed model, we test it on five commonly used benchmark datasets: Set5, Set14, BSDS100, Urban100 and Manga109, each of which has different characteristics.

In our experiment, VGG13 [20] is deployed as the backbone of our discriminator, as other works [11, 28] did. Since max pooling operation may lose some information during feed-forward process, we instead use the convolution kernel with stride set to 2 to downsample the feature maps. We empirically choose the hyper-parameters $\lambda = 5 \times 10^{-3}$ and $\eta = 1 \times 10^{-2}$, the the initial learning rate is 10^{-4}. Using Adam optimizer, the models are trained with a batch size of 16 with the learning rate reduing to half every 200k iterations.

4.2. Evaluation metrics

In recent years, SR algorithms have gradually developed into two directions: one is to obtain higher restoration accuracy measured by PSNR [13, 36, 12] which calculates the pixel-wise difference between SR image and groundtruth:

$$\mathrm{PSNR} = 10 \cdot \log_{10}\frac{{\mathrm{MAX_I}}^2}{\mathrm{MSE}}, \tag{11}$$

where the higher is the better. The other is to measure perceptual quality of reconstructed images, among which the perceptual index [28] is the most commonly used metric:

$$\mathrm{PI} = \frac{1}{2}\left((10 - \mathrm{Ma}) + \mathrm{NIQE}\right), \tag{12}$$

which combines the no-reference image quality measures of Ma score [14] and NIQE [16], and the lower is the better.

However, the two objective metrics generally grow in an opposite way. The lower the PI value, the higher the PSNR, and vice versa. In this paper, for better visual quality, we employ PI as the main criteria and PSNR as an auxiliary.

4.3. Ablation study

In this section, we will verify the effects of our proposed hierarchical feature extraction module (HFEM) and hierarchical guided reconstruction module (HGRM). We train several models with different configurations and test them on Set14, as listed in Table 1. Each row represents a model with its settings.

First of all, we set up a baseline (*model 1*) which only contains one branch feature extractor ($B = 1$) with filter

Models	HFEM	HGRM	PI / PSNR	Training time (days)
1	$B = 1\,(k = 3)$	$T = 1\,(f = 4)$	3.109 / 26.344	2.1
2	$B = 2\,(k = 3, 5)$	$T = 1\,(f = 4)$	3.060 / 25.609	2.3
3	$B = 3\,(k = 3, 5, 7)$	$T = 1\,(f = 4)$	2.810 / 25.263	2.8
4	$B = 4\,(k = 3, 5, 7, 9)$	$T = 1\,(f = 4)$	2.796 / 25.436	3.8
5	$B = 3\,(k = 3, 5, 7)$	$T = 2\,(f = 2, 4)$	2.897 / 26.239	2.9
6	$B = 3\,(k = 3, 5, 7)$	$T = 3\,(f = 1, 2, 4)$	2.903 / 26.234	2.9

Table 1. The effects of different components of our HSRGAN. Each row represents a model with its configurations of HFEM and HGRM. k means the kernel size of each branch in HFEM and f denotes the corresponding upscale factor of each branch in HGRM. The quantitative results (perceptual index and PSNR) and training time is listed on the right.

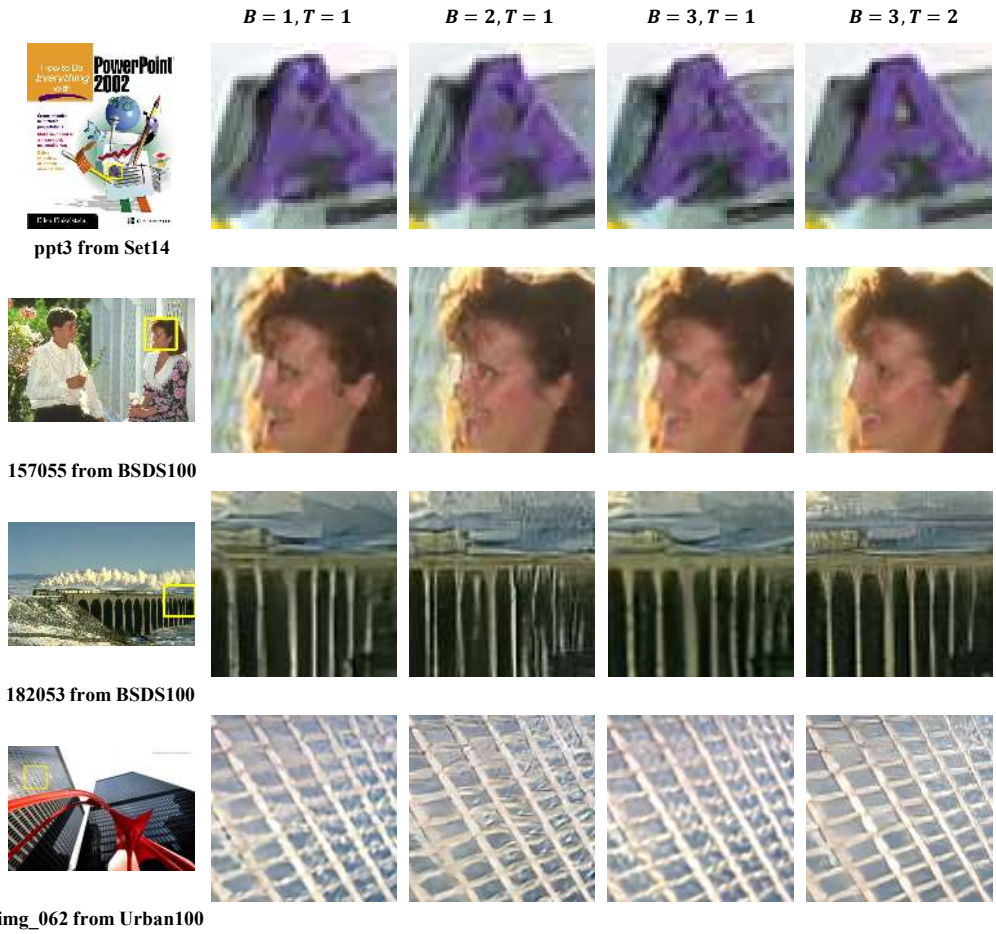

Figure 4. Overall visual comparisons for showing the effects of each component in our models. Each column represents a model with its configurations in Table 1.

size of 3 and an upscale module without an intermediate branch ($T = 1$). From Table 1 we can see that *model 1* achieves PI $= 3.109$ and PSNR $= 26.344$. Then we employ our HFEM and add branches as *model 2* to *model 4*. Results from *model 2* to *model 4* verify the effectiveness of the hierarchical features extracted from our HFEM. Specifically, *model 3* significantly optimizes PI to 2.810 and *model 4* further reduces PI to 2.796 while costs more than 1 day in training, since the computing complexity caused by kernel size of 9. In the meantime, the pixel-wise accuracy reflected by PSNR deteriorates about 1dB.

Furthermore, based on *model 3*, which balances the quantitative results and computing efficiency, we add intermediate branches to verify the functionality of our proposed HGRM. Specifically, *model 5* contains an intermediate branch of upscale factor of 2 ($T = 2$) and *model 6* contains one more branch of upscale factor of 1 ($T = 3$), which guides the output of HFEM to estimate the original

Dataset	Bicubic	EDSR [13]	RCAN [36]	SRFBN [12]	SRGAN [11]	SFTGAN [27]	NatSR [21]	ESRGAN [28]	HSRGAN (Ours)
Set5	7.369	5.962	5.958	5.937	**3.536**	3.759	4.165	3.755	3.688
Set14	7.027	5.285	5.246	5.403	2.948	2.906	3.109	2.926	**2.897**
BSDS100	7.003	5.258	5.130	5.352	2.381	2.377	2.780	**2.313**	2.406
Urban100	6.944	4.989	4.987	5.138	3.495	3.614	3.652	3.635	**3.369**
Manga109	6.764	4.718	4.760	4.871	3.370	3.308	3.463	3.416	**3.295**

Table 2. Quantitative evaluation of state-of-the-art SR approaches on datasets Set5, Set14, BSDS100, Urban100, and Manga109 measured by perceptual index. Best and second best results are **highlighted** and underlined.

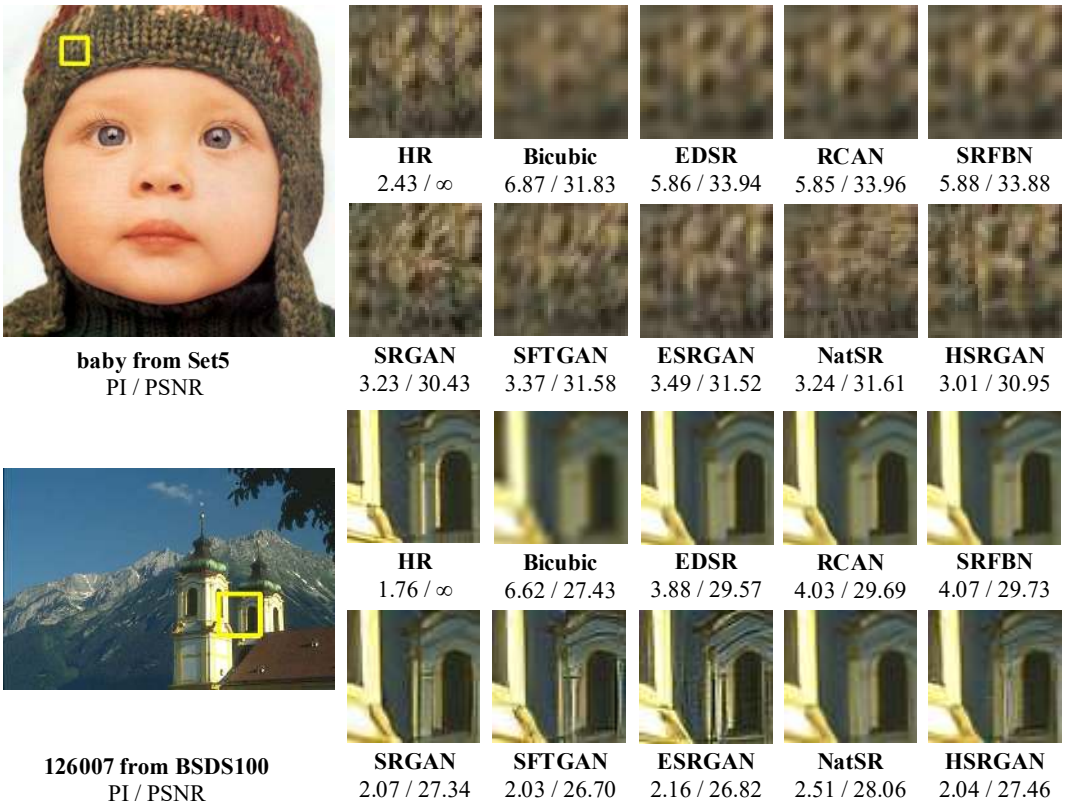

Figure 5. Qualitative results of our HSRGAN and state-of-the-art methods. HSRGAN generates more realistic textures and less artifacts.

LR image. Experiments indicate that with the help of intermediate guided supervision, our model can generate higher quality images with a raise of PSNR by nearly 1dB. Nevertheless, *model 6* achieves almost the same quantitative results as *model 5* in PI and PSNR, which implies that the adding the intermediate branch of original resolution does not further improve the performance. Comparing *model 1* with *model 5* or *model 6*, it demonstrates that our hierarchical generative adversarial networks composed of HFEM and HGRM can reduce the PI while maintains PSNR at the same time.

To compare the visual quality of different models, we test *model 1*, *model 2*, *model 3* and *model 5* on BSDS100 and Urban100. And the overall visual comparisons are illustrated in Figure 4. Comparing *model 1*, *model 2* and *model 3*, we can see the significant improvement on the image clarity. *model 5* recovers much clearer edges of letter *A* in image "ppt3" and more natural facial details in image "157055". It demonstrates that the hierarchical features captured by multi-branch network help with the sharpness. From *model 3* and *model 5*, we verify that the HGRM can reduce the artifacts generated by HFEM, like arches in image "182053" and windows in image "img_062", and make the image more natural. The main reason lies in that the structure information can be reconstructed by intermediate supervision of HGRM.

4.4. Comparison with the state-of-the-art

In this section, we employ Bicubic, EDSR [13], RCAN [36], SRFBN [12], SRGAN [11], SFTGAN [27], NatSR

[21], ESRGAN [13] as our comparison methods. We retrained these models with their published codes and run them on the test datasets. Table 2 lists the quantitative results. The methods can be divided into 3 categories, where Bicubic is a baseline for the others. Compared to distortion-oriented methods, such as EDSR, RCAN and SRFBN, GAN-based or perception-oriented methods show significant advantages in perceptual index, which indicates that the methods generate clear edges of images to some extent. Among all the GAN-based methods, our HSRGAN outperforms the other methods on Set14, Urban100, Manga109 datasets and achieves the second best on Set5 dataset, which is comprehensively the best quantitative results.

We also show the qualitative results of various methods in Figure 5. As we can see, the distortion-oriented methods produce over-smoothing images, while recent GAN-based methods outperform in both sharpness and details. Taking image "baby" as an example, EDSR, RCAN and SRFBN obtain more faithful results of the woolen hat. SRGAN, SFTGAN and ESRGAN reconstruct the clear textures but fail to reproduce the natural shape of wool. NatSR generates sensible wool textures but it is a little blurry. Our HSRGAN is capable of generating clear details and realistic wool textures at the same time. Another problem of GAN-based methods is that they sometimes add undesired noise into the final SR images. In image "126007", SFTGAN and ESRGAN recover the abrupt surface of the tower especially the left part and top part of the window. SRGAN produces a more natural tower surface but there still exist rigid artifacts on the left part. Our HSRGAN can get rid of the unpleasant artifacts while maintaining enough details and generate clearer image than NatSR. More qualitative comparison can be found in our *supplementary material*.

4.5. Subjective assessments

However, perceptual index does not fully reflect the visual quality of the image. The lower perceptual index does not always guarantee a better visual quality. [1] points out that perceptual index is correlated with the human-opinion-scores on a coarse scale, but it is not always well-correlated with these scores on a finer scale. To provide a better reference standard for visual quality assessment, we use the mean opinion score (MOS) to quantify our performance.

Specifically, we ask 30 raters to assign a score from 1 (bad quality) to 5 (excellent quality) to the super-resolved images of BSDS100. The raters rate 10 randomly-shuffled versions of super-resolved each image on BSDS100 recovered by Bicubic, EDSR [13], RCAN [36], SRFBN [12], SRGAN [11], SFTGAN [27], NatSR [21], ESRGAN [13], our HSRGAN and the original HR image. Each rater thus rates 1000 instances (100 images $\times$ 10 versions) and each instance is rated 30 times. The results of distribution of MOS scores are illustrated as Figure 6. It is worth noting that a method rated less than 10 times will not be shown in the figure for better visualization.

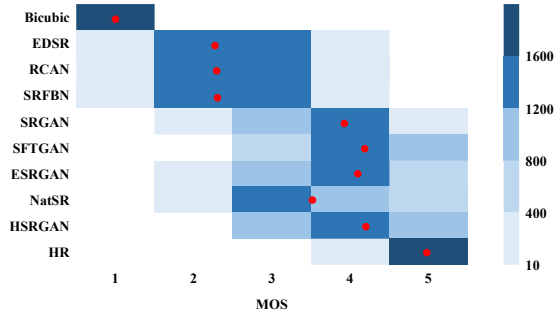

Figure 6. Color-coded distribution of MOS scores on BSDS100. For each method 3000 samples (100 images $\times$ 30 raters) are assessed. Mean shown as red marker and the darker the color, the more times the score of the method. The distortion-oriented methods focus on 2 or 3 points, while the most perception-oriented methods focus on 4 points.

As we can see that raters very consistently rate bicubic interpolated test images as 1 and the original HR images as 5. The scores of EDSR, RCAN and SRFBN mainly lie in 2 or 3 and are extremely close since they focus on minimizing pixel-wise distortion, such as MAE or MSE, resulting in producing the mean result from intent multiple HR counterparts, which is not helpful to improve the subjective visual quality. For the rest GAN-based methods, the scores range from 2 to 5. And our HSRGAN slightly outperforms SFTGAN, NarSR, ESRGAN and SRGAN.

5. Conclusions

We proposed hierarchical generative adversarial networks (HSRGAN) for the SISR problem. Specifically, the hierarchical feature extraction module (HFEM) extracts the hierarchical features in multiple receptive fields, concentrating on both local texture and global semantics. In addition, we proposed a hierarchical guided reconstruction module (HGRM). It reconstructs the SR image by adding intermediate supervision branches in a progressive manner. Our HSRGAN can generate more sensible structural textures than directly upsampling a LR image to a large magnification. Extensive experiments on 5 common datasets show that our method achieves state-of-the-art performance in terms of both quantitative metrics and visual quality.

Acknowledgements

This work was supported by The National Key Research and Development Plan of China (2020AAA0103502), National Natural Science Foundation of China (62022009 and 61872021), Beijing Nova Program of Science and Technology (Z191100001119050), and the Academic Excellence Foundation of BUAA for PhD Students.

References

[1] Yochai Blau, Roey Mechrez, Radu Timofte, Tomer Michaeli, and Lihi Zelnik-Manor. The 2018 pirm challenge on perceptual image super-resolution. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 0–0, 2018.

[2] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4690–4699, 2019.

[3] Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang. Learning a deep convolutional network for image super-resolution. In *European conference on computer vision*, pages 184–199. Springer, 2014.

[4] Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang. Image super-resolution using deep convolutional networks. *IEEE transactions on pattern analysis and machine intelligence*, 38(2):295–307, 2015.

[5] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014.

[6] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[7] Andrew Howard, Mark Sandler, Grace Chu, Liang-Chieh Chen, Bo Chen, Mingxing Tan, Weijun Wang, Yukun Zhu, Ruoming Pang, Vijay Vasudevan, et al. Searching for mobilenetv3. *arXiv preprint arXiv:1905.02244*, 2019.

[8] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7132–7141, 2018.

[9] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708, 2017.

[10] Alexia Jolicoeur-Martineau. The relativistic discriminator: a key element missing from standard gan. *arXiv preprint arXiv:1807.00734*, 2018.

[11] Christian Ledig, Lucas Theis, Ferenc Huszár, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew Aitken, Alykhan Tejani, Johannes Totz, Zehan 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*, pages 4681–4690, 2017.

[12] Zhen Li, Jinglei Yang, Zheng Liu, Xiaomin Yang, Gwanggil Jeon, and Wei Wu. Feedback network for image super-resolution. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3867–3876, 2019.

[13] Bee Lim, Sanghyun Son, Heewon Kim, Seungjun Nah, and Kyoung Mu Lee. Enhanced deep residual networks for single image super-resolution. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 136–144, 2017.

[14] Chao Ma, Chih-Yuan Yang, Xiaokang Yang, and Ming-Hsuan Yang. Learning a no-reference quality metric for single-image super-resolution. *Computer Vision and Image Understanding*, 158:1–16, 2017.

[15] Yuqing Ma, Xianglong Liu, Shihao Bai, Lei Wang, Dailan He, and Aishan Liu. Coarse-to-fine image inpainting via region-wise convolutions and non-local correlation. In *Proceedings of the 28th International Joint Conference on Artificial Intelligence*, pages 3123–3129. AAAI Press, 2019.

[16] Anish Mittal, Rajiv Soundararajan, and Alan C Bovik. Making a "completely blind" image quality analyzer. *IEEE Signal Processing Letters*, 20(3):209–212, 2012.

[17] Seong-Jin Park, Hyeongseok Son, Sunghyun Cho, Ki-Sang Hong, and Seungyong Lee. Srfeat: Single image super-resolution with feature discrimination. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 439–455, 2018.

[18] Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*, 2015.

[19] Mehdi SM Sajjadi, Bernhard Scholkopf, and Michael Hirsch. Enhancenet: Single image super-resolution through automated texture synthesis. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4491–4500, 2017.

[20] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[21] Jae Woong Soh, Gu Yong Park, Junho Jo, and Nam Ik Cho. Natural and realistic single image super-resolution with explicit natural manifold discrimination. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[22] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–9, 2015.

[23] Ying Tai, Jian Yang, Xiaoming Liu, and Chunyan Xu. Memnet: A persistent memory network for image restoration. In *Proceedings of the IEEE international conference on computer vision*, pages 4539–4547, 2017.

[24] Radu Timofte, Eirikur Agustsson, Luc Van Gool, Ming-Hsuan Yang, and Lei 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*, pages 114–125, 2017.

[25] Radu Timofte, Vincent De Smet, and Luc Van Gool. Anchored neighborhood regression for fast example-based super-resolution. In *Proceedings of the IEEE international conference on computer vision*, pages 1920–1927, 2013.

[26] Radu Timofte, Vincent De Smet, and Luc Van Gool. A+: Adjusted anchored neighborhood regression for fast super-resolution. In *Asian conference on computer vision*, pages 111–126. Springer, 2014.

[27] Xintao Wang, Ke Yu, Chao Dong, and Chen Change Loy. Recovering realistic texture in image super-resolution by deep spatial feature transform. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 606–615, 2018.

[28] Xintao Wang, Ke Yu, Shixiang Wu, Jinjin Gu, Yihao Liu, Chao Dong, Yu Qiao, and Chen Change Loy. Esrgan: Enhanced super-resolution generative adversarial networks. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 0–0, 2018.

[29] Yi Wang, Xin Tao, Xiaojuan Qi, Xiaoyong Shen, and Jiaya Jia. Image inpainting via generative multi-column convolutional neural networks. In *Advances in Neural Information Processing Systems*, pages 331–340, 2018.

[30] Zhou Wang, Alan C Bovik, Hamid R Sheikh, Eero P Simoncelli, et al. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004.

[31] Jianchao Yang, John Wright, Thomas Huang, and Yi Ma. Image super-resolution as sparse representation of raw image patches. In *2008 IEEE conference on computer vision and pattern recognition*, pages 1–8. Citeseer, 2008.

[32] Jianchao Yang, John Wright, Thomas S Huang, and Yi Ma. Image super-resolution via sparse representation. *IEEE transactions on image processing*, 19(11):2861–2873, 2010.

[33] Fisher Yu and Vladlen Koltun. Multi-scale context aggregation by dilated convolutions. *arXiv preprint arXiv:1511.07122*, 2015.

[34] Kaibing Zhang, Xinbo Gao, Dacheng Tao, and Xuelong Li. Single image super-resolution with non-local means and steering kernel regression. *IEEE Transactions on Image Processing*, 21(11):4544–4556, 2012.

[35] Lei Zhang and Xiaolin Wu. An edge-guided image interpolation algorithm via directional filtering and data fusion. *IEEE transactions on Image Processing*, 15(8):2226–2238, 2006.

[36] Yulun Zhang, Kunpeng Li, Kai Li, Lichen Wang, Bineng Zhong, and Yun Fu. Image super-resolution using very deep residual channel attention networks. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 286–301, 2018.

[37] Yulun Zhang, Yapeng Tian, Yu Kong, Bineng Zhong, and Yun Fu. Residual dense network for image super-resolution. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2472–2481, 2018.