Yue HJ, Shen S, Yang JY *et al.* Reference image guided super-resolution via progressive channel attention networks. JOURNAL OF COMPUTER SCIENCE AND TECHNOLOGY 35(3): 551–563 May 2020. DOI 10.1007/s11390-020-0270-3

Reference Image Guided Super-Resolution via Progressive Channel Attention Networks

Huan-Jing Yue[1], *Member*, *IEEE*, Sheng Shen[1], Jing-Yu Yang[1,*], *Senior Member*, *IEEE*, Hao-Feng Hu[2], and Yan-Fang Chen[1]

[1] *School of Electrical and Information Engineering, Tianjin University, Tianjin 300072, China*
[2] *School of Precision Instrument and Opto-Electronics Engineering, Tianjin University, Tianjin 300072, China*

E-mail: {huanjing.yue, codyshens, yjy, haofeng_hu, cyf}@tju.edu.cn

Received January 18, 2020; revised April 2, 2020.

Abstract In recent years, the convolutional neural networks (CNNs) for single image super-resolution (SISR) are becoming more and more complex, and it is more challenging to improve the SISR performance. In contrast, the reference image guided super-resolution (RefSR) is an effective strategy to boost the SR (super-resolution) performance. In RefSR, the introduced high-resolution (HR) references can facilitate the high-frequency residual prediction process. According to the best of our knowledge, the existing CNN-based RefSR methods treat the features from the references and the low-resolution (LR) input equally by simply concatenating them together. However, the HR references and the LR inputs contribute differently to the final SR results. Therefore, we propose a progressive channel attention network (PCANet) for RefSR. There are two technical contributions in this paper. First, we propose a novel channel attention module (CAM), which estimates the channel weighting parameter by weightedly averaging the spatial features instead of using global averaging. Second, considering that the residual prediction process can be improved when the LR input is enriched with more details, we perform super-resolution progressively, which can take advantage of the reference images in multi-scales. Extensive quantitative and qualitative evaluations on three benchmark datasets, which represent three typical scenarios for RefSR, demonstrate that our method is superior to the state-of-the-art SISR and RefSR methods in terms of PSNR (Peak Signal-to-Noise Ratio) and SSIM (Structural Similarity).

Keywords reference-based super resolution, channel attention, progressive channel attention network (PCANet)

1 Introduction

Image super-resolution (SR) aims to predict a high-resolution (HR) image from a low-resolution (LR) observation. One LR image can correspond to many HR observations, which makes the LR to HR mapping process much difficult. Traditional interpolation-based methods, such as bilinear and bicubic interpolation, utilize the local statistics of the LR image to estimate the HR image, which cannot produce realistic details. Later on, learning-based approaches, especially deep learning based approaches, are proposed to estimate the mapping between LR and HR, showing great success in the SR field [1–9]. However, most existing methods still suffer from smooth results at upscaling factor 4x or larger, especially when there are very rich details in the original HR image but they are lost in the corresponding LR image. Meanwhile, with the explosive growth of Internet images, storage capacity, and the multiview capturing devices, it is likely to find similar images from the Internet, personal albums, and multiview imaging systems. Therefore, the reference image guided SR (RefSR), which hallucinates the HR image with the guidance of similar reference images, is proposed [10–12]. RefSR utilizes the abundant detail textures from HR reference images to make up for the

Regular Paper
Special Section of CVM 2020
This work was supported in part by the National Natural Science Foundation of China under Grant Nos. 61672378, 61771339, and 61520106002.
*Corresponding Author

lost details of LR images. The first RefSR [10] utilizes retrieved reference images to assist reconstruction via patch matching and patch fusion. Hereafter, the sparse learning based methods, which estimate the parameters via sparse coding, are proposed to improve the fusion process [13, 14]. However, the hand-crafted priors cannot effectively deal with different similarity levels. Recently, the deep learning based RefSR [11, 12, 15, 16] have been proposed. The work in [15, 16] explores both the internal and the external correlations in the pixel space to upsample the LR input. Meanwhile, Zhang *et al.* proposed to enrich the HR details by transferring similar textures from the reference images [11] via matching in the neural space. These methods have achieved superior performance compared with the single image based SR (SISR).

However, the above deep learning based RefSR methods just simply merge the reference feature and the LR feature together, without considering the relevance levels of different channels. To address this issue, we propose a novel RefSR method by incorporating the channel attention module in the feature fusion process. Our contributions are summarized as follows.

• We propose a novel channel attention module, which estimates the channel weighting parameter by adaptively merging the spatial features instead of using the average pooling. With the proposed channel attention module, we can assign larger weight to the reference block which has a larger similarity to the LR input.

• Considering that the residual prediction can be improved when the LR input is enriched with more details, we propose to upsample the LR image progressively. At each upsampling level, the high-frequency residual details are predicted by using the information from the reference images, the original LR input, and the SR result from the previous upsampling level.

• We evaluate the proposed method on three benchmark datasets and our method achieves the best SR results in both subjective and objective measurements. The ablation study demonstrates that the proposed channel attention module and the progressive reconstruction strategy greatly improve the RefSR performance.

2 Related Work

In this section, we give a brief review of deep learning based single image SR, RefSR, and the attention module in convolutional neural networks.

2.1 Deep Learning Based Single Image Super Resolution

Traditional SISR methods mainly utilize example-based approaches to learn a mapping between LR and HR patches [17–22]. The hand-crafted mapping strategy limits the SR performance. In recent years, deep learning based SISR has presented significant advantages in both quantitative and qualitative results compared with traditional algorithms. Dong *et al.* first introduced CNN into the SR field by building a three-layer convolutional network SRCNN [1], which achieved outstanding performance compared with previous work. VDSR [3] and DRCN [23] were proposed by Kim *et al.* They extended CNN to 20 layers and effectively improved the SR performance by using residual learning.

However, these methods need to first interpolate the input LR to the target scale, which is computationally intensive. Therefore later researchers proposed to perform convolution operations on the initial LR image and used the network to upsample it to the target size at the last layer. Shi *et al.* [24] introduced a sub-pixel convolutional layer at the end of the network to replace traditional interpolation, which is more efficient for reconstruction. LapSRN [4] was proposed to progressively upsample the LR images to its desired resolution. Ledig *et al.* introduced SRResnet [25] based on ResNet [26] for image super resolution. They also proposed SRGAN [25], which utilizes GAN [27] and perceptual loss [28] to improve the visual quality of the SR results. EDSR [6] proposed by Lee *et al.*, stood on the shoulders of SRResnet, and used residual blocks to restore the HR image. Recently, RCAN [8] has been proposed, which introduces channel attention module in the residual blocks, and achieves state-of-the-art SISR performance.

From SRCNN to RCAN, the SR performance is greatly improved. However, we also observe that the network structures are becoming more and more complex, while the improvement is becoming smaller. In this paper, we utilize reference images to boost the SR performance.

2.2 Reference-Based Super Resolution

Unlike SISR, where only one single LR image is used for input, RefSR utilizes similar HR reference images to guide the SR process. Landmark [10] proposed by Yue *et al.*, is the first work for RefSR, which super resolves the LR input by traditional patch matching and blending scheme. Then the work in [13] introduces sparse coding to improve the parameter estimating process in

Landmark. The two methods are both based on handcrafted priors, which limit their SR performance.

Recently, the CNN-based RefSR methods have emerged. Yue *et al.*[16] and Yang *et al.*[15] proposed CNN-based RefSR methods which explore both the internal and the external correlations. Crossnet[12] was proposed to deal with light field image SR, and the HR reference images have only small displacements compared with the LR input. This enables CrossNet to utilize optical flow to align the HR reference with the LR input. However, this makes it cannot deal with reference images with large displacements with the LR input. Zhang *et al.*[11] proposed SRNTT to adaptively transfer textures from reference images according to their textural similarity, which can take advantage of reference images with different similarity levels. However, the neural space matching will degrade the SR performance when the reference images are similar to the LR input.

Besides, all the CNN-based RefSR methods do not introduce attention modules to adaptively utilize the features from LR input and reference images. In this paper, we propose a channel attention module to assign larger weights to the reference block which has a larger similarity to the LR input.

2.3 Attention Mechanisms

The purpose of the attention mechanism in neural networks is to recalibrate the feature response for the most beneficial and important part of the previous layer input. Recently, some work has focused on integrating attention modules into a series of tasks, such as image classification[29], image generation[30], and image restoration[7,8,31]. By studying the inter-dependence between convolutional feature channels in the network, Hu *et al.*[29] introduced a channel attention mechanism called squeeze and excitation (SE) block to adaptively recalibrate the channel feature response for image classification. Inspired by SE networks, Zhang *et al.*[8] proposed RCAN which combines channel attention with residual blocks to form a very deep residual network, achieving the state-of-the-art performance of SISR. Furthermore, CBAM[32] explores both channel-wise and spatial-wise relationship of feature maps via CA and SA modules. However, the above-mentioned attention methods all utilize global average/max pooling to get channel/spatial wise statistics information. Different from them, we propose a weighted average pooling method for channel attention.

3 Proposed Method

In this section, we first give an overview of the proposed method and then present the details for each module.

3.1 Framework Overview

Given an LR image I^L as input, we aim to recover the HR image $\hat{I}^H$ from I^L with the guidance of similar HR reference images $\{I_1^R, I_2^R, ..., I_n^R\}$①. Since the reference images are captured with different viewpoints, focal lengths, and illuminations, and may contain different objects, it is unreasonable to directly concatenate the LR input and reference image together to infer the HR details. Therefore, we perform the LR to the HR mapping at the patch level. To improve the patch matching accuracy, we first align the reference images with the LR input. Then we retrieve similar patches from the aligned reference images $\{\tilde{I}_1^R, \tilde{I}_2^R, ..., \tilde{I}_n^R\}$. Hereafter, we feed the matched HR patches and the original LR patch into the proposed progressive channel attention network (PCANet) to infer the HR details. After recovering all the HR patches, they are blended together via averaging in the overlapped regions to produce the predicted HR image $\hat{I}^H$.

In the following, we give details of patch matching and the proposed PCANet.

3.2 Reference Image Alignment and Patch Matching

3.2.1 Reference Image Alignment

Directly searching for similar patches from the reference images not only involves huge computing complexity but also may miss the most similar patches since there is deformation between the reference image and the LR input. Therefore, we propose to first align the reference images with the LR input according to their matched feature points. For a given reference image I^R and the bicubic interpolation version of I^L, denoted as $(I^L)^{\uparrow}$, whose size is our target size, we first extract their SIFT[33] feature points. Then, we perform the feature matching using the matching criteria proposed in [33] to find the matched points between I^R and $(I^L)^{\uparrow}$. Hereafter, we utilize the matched points to regress a

① These reference images can be obtained via image retrieval from cloud database, photo albums, videos, or multi-view imaging systems, which is out of the scope of this paper.

homography transform matrix using the RANSAC [34] algorithm. Finally, we get the aligned reference image $\tilde{I}^R$ by transforming the reference image I^R with the corresponding homography matrix to make $\tilde{I}^R$ have a similar scale and viewpoint to $(I^L)^{\uparrow}$. For more information, please refer to [10].

3.2.2 Patch Matching

For the LR patch $(P^L)^{\uparrow}$ in $(I^L)^{\uparrow}$, we aim to find its matched high frequency (HF) details Q^{HF} from $\{\tilde{I}_1^R, \tilde{I}_2^R, ..., \tilde{I}_n^R\}$. Considering that the super-resolving process will keep the low frequency information, we construct a new HR reference using the original low frequency (LF) information and the HF information from the reference image, namely $P^R = (P^L)^{\uparrow} + Q^{HF}$.

Therefore, we decompose the reference image into HF and LF parts. The LF part of $\tilde{I}^R$ is obtained by first downsampling it and then upsampling it to the original resolution, denoted by $(\tilde{I}^R)^{\downarrow\uparrow}$. The dowmsampling and upsampling ratio is the same as our targeted upsampling ratio for the LR input I^L. In this way, each reference image can be decomposed as $\tilde{I}^R = (\tilde{I}^R)^{\downarrow\uparrow} + (\tilde{I}^R)^{HF}$. Then, for each patch $(P^L)^{\uparrow}$ in $(I^L)^{\uparrow}$, we search for its matched patches Q^L from $\{(\tilde{I}_1^R)^{\downarrow\uparrow}, (\tilde{I}_2^R)^{\downarrow\uparrow}, ..., (\tilde{I}_n^R)^{\downarrow\uparrow}\}$. Considering $(\tilde{I}_i^R)^{\downarrow\uparrow}$ is aligned with $(I^L)^{\uparrow}$, for a patch $(P^L)^{\uparrow}$ of size $m \times m$ centered at position(x, y), we constrain the search window to be a region of size $2m \times 2m$ centered at (x, y), denoted as W^P. The best matched k patches are obtained by minimizing the Euclidean distance, i.e.,

$$D((P^L)^{\uparrow}, Q_j^L) = \|(P^L)^{\uparrow} - Q_j^L\|_2^2,$$

where Q_j^L is the candidate patch, densely sampled from the region W^P of $\{(\tilde{I}_1^R)^{\downarrow\uparrow}, (\tilde{I}_2^R)^{\downarrow\uparrow}, ..., (\tilde{I}_n^R)^{\downarrow\uparrow}\}$. After obtaining the best matched k patches $\{Q_1^L, Q_2^L, ..., Q_k^L\}$, we extract their corresponding HF patches $\{Q_1^{HF}, Q_2^{HF}, ..., Q_k^{HF}\}$ from $\{(\tilde{I}_1^R)^{HF}, (\tilde{I}_2^R)^{HF}, ..., (\tilde{I}_n^R)^{HF}\}$. Finally, we construct the reference patches by

$$P_i^R = (P^L)^{\uparrow} + Q_i^{HF}, i \in \{1, 2, ..., k\},$$

for the input $(P^L)^{\uparrow}$.

Note that, all the operations are performed in the brightness channel and the DC component is removed in patch matching since there are usually differences in illumination and chromaticity between the LR input and the reference images. In our experiments we set the patch size $m \times m$ to be 20×20. On the one hand, an LR patch with a larger size includes more structure information, but it increases the matching error if there is no exactly matched patch. On the other hand, a patch with a smaller size is more flexible in finding matched patches, but it will limit the receptive field of the following PCANet. In our experiments, we find 20×20 is suitable for most cases.

3.3 Proposed PCANet

The proposed PCANet progressively upsamples the LR input to a large resolution. For a given upsampling scale S, our network contains $\log_2 S$ levels. For example, our network has two levels when the upsampling scale S is 4. Fig.1 presents our PCANet at 4x upsampling. The input of our network is the original LR patch P^L from I^L and the reconstructed HR reference patches $\{P_1^R, P_2^R, ..., P_k^R\}$. Then, we perform the first 2x upsampling using bicubic interpolation on P^L. Correspondingly, the HR reference patches are downsampled by the scaling factor of 2 to match the resolution of $(P^L)^{\uparrow 2x}$. Then, $(P^L)^{\uparrow 2x}$ and $\{(P_1^R)^{\downarrow \frac{1}{2}x}, (P_2^R)^{\downarrow \frac{1}{2}x}, ..., (P_k^R)^{\downarrow \frac{1}{2}x}\}$ are concatenated together to go through the first-level HF detail prediction network. Hereafter, the SR result of the first-level, i.e., $\hat{P}^{H_1}$ is further upsampled by a factor of 2 using the sub-pixel convolution [24]. This process is denoted by $f_{\text{up}}(\hat{P}^{H_1})$. Then, we concatenate $f_{\text{up}}(\hat{P}^{H_1})$, $(P^L)^{\uparrow 4x}$, and $\{P_1^R, P_2^R, ..., P_k^R\}$, and feed them into the second-level HF detail prediction network. Finally, we get the SR result at the second level, denoted by $\hat{P}^{H_2}$. In the following, we give details of the proposed PCANet.

3.3.1 Network Structure

We only present the network details of the first-level upsampling since the network structures of the following levels are the same as those of the first level. As shown in Fig.1, the PCANet is constructed by the channel attention module (CAM) and the residual channel attention module (RCAM). In the following, we give details about the two modules.

1) *Channel Attention Module.* Assume that F_{in} is the input feature of CAM and F_{out} is its output feature. The relation between F_{in} and F_{out} can be formulated as:

$$F_{\text{out}} = f_{\text{sc}}(F_{\text{in}}, f_{\text{ex}}(f_{\text{sq}}(F_{\text{in}}))),$$

where f_{sq}, f_{ex}, and f_{sc} represent the squeeze, excitation, and scaling processes, respectively. In the previous work [8, 29], the squeeze process simply utilizes global average pooling to extract statistics information of the channels. The excitation process excites the information through the sigmoid layer to produce channel weights, and the scaling process multiplies the feature

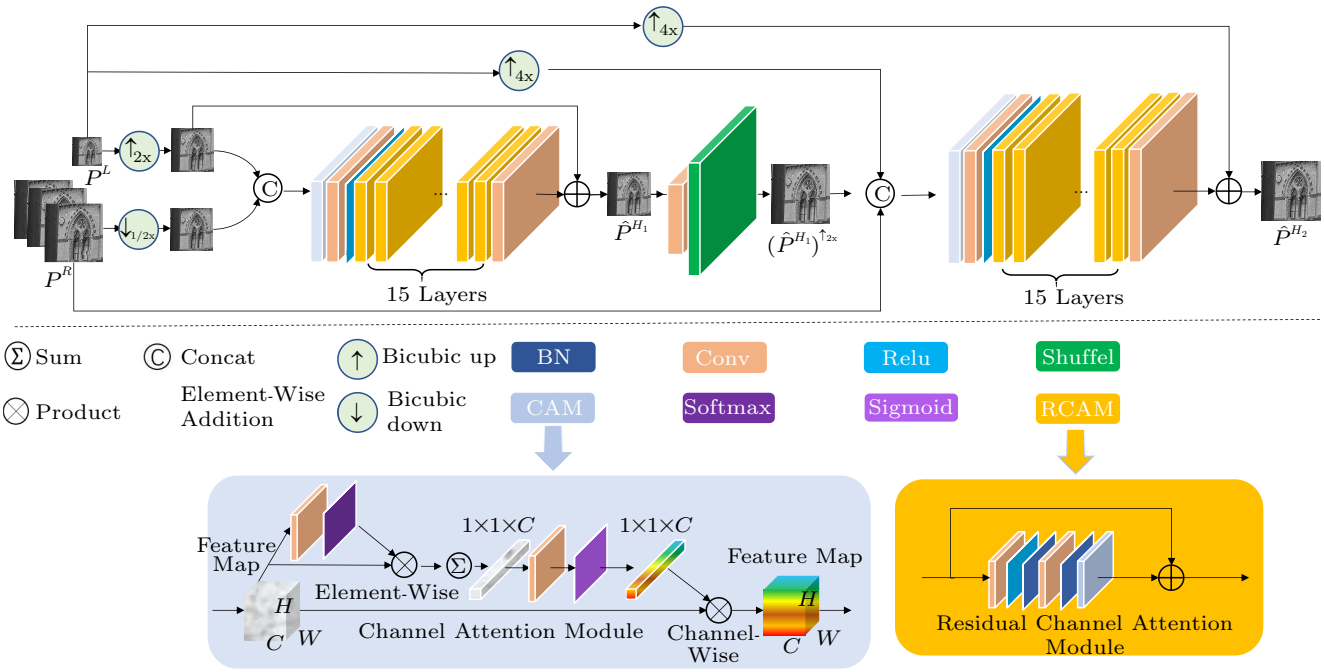


Fig.1. Framework of the proposed PCANet. P^L and P^R are the LR patch and corresponding reference patches, respectively. $\hat{P}^{H_i}$ denotes the output of the i-th level. C is the number of channels

maps F_{in} with the channel weights. In this way, the features which contribute more to the final result are multiplied with higher weights. This strategy works well in high-level vision problems, such as image classification and semantic segmentation. However, in RefSR, it is coarse to estimate the channel weights using the statistic information obtained by global average pooling since the features in different spatial positions may have different contributions to the final result. Therefore, we propose to give different weights to the pixels/features in different spatial positions in the pooling process of the squeeze process. The proposed squeeze process is defined as:

$$\boldsymbol{\gamma} = f_{\text{sq}}(F_{\text{in}}) = \frac{1}{HW}\sum_{h,w} F_{\text{in}} \times f_{\text{s}}(f_{\text{c}}(F_{\text{in}})),$$

where f_{c} denotes the convolution operation to get the weight for each point, and f_{s} is the Softmax function to normalize weights in the same channel. $\times$ represents pixel-wise multiplication. The sum operation is performed along the height (H) and the width (W) dimensions. In this way, we extend the original global average pooling to weighted average pooling.

Our excitation process is denoted by

$$f_{\text{ex}}(\boldsymbol{\gamma}) = \sigma(\mathcal{W}(\boldsymbol{\gamma})),$$

where $\mathcal{W}$ is the fully connection layer (realized by 1×1 convolution) and σ is the sigmoid layer. This process produces the weights for each channel. Hereafter, each channel of the feature map F_{in}, i.e., F_{in}^c, is multiplied by the scalar weight γ_c, namely $F_{\text{out}}^c = \gamma_c F_{\text{in}}^c$. The framework of the proposed CAM module is depicted in Fig.1.

2) *Residual Channel Attention Module.* The proposed RCAM module, as depicted in Fig.1, is mainly constructed by two convolution layers and one CAM layer. Namely, we integrate the CAM module into the original residual block to make it give different weights to different feature channels.

All the convolutional filters depicted in Fig.1 are of size 3×3 and the channel number is 96 except for the output layer, whose channel number is 1 for gray images.

3.3.2 Loss Function

We optimize the proposed PCANet via minimizing the L_2 distance between the recovered HR patch $\hat{P}^H$ and the corresponding ground-truth HR patch P^H at each level. The objective function for training is formulated as:

$$\mathcal{L}(P^H, \hat{P}^H; \Theta) = \frac{1}{N}\sum_{i=1}^{N}\sum_{s=1}^{\log_2 S} \|P_i^{H_s} - \hat{P}_i^{H_s}\|_2^2,$$

where Θ denotes the network parameter, N is the number of training pairs in each minibatch, and s is the upsampling level.

3.3.3 Training Details

All RGB low-resolution images are generated by downsampling the corresponding HR images using bicubic interpolation. Since humans are more sensitive to the luminance changes, we only perform super-resolution on the luminance channel (Y) and the chrominance channels (UV) are directly upsampled via bicubic upsampling.

We utilize the training set released in the work of [16], which contains 15 groups of landmark images. Each target image is split into overlapped 20×20 patches at the step size of 4. We totally extract nearly 700 000 patches for training. During training, we perform data augmentation via flips and rotations.

In the training phase, the batch size is set to 128. We train our networks with Adam [35] optimizer, where the parameters of the optimizer are set as $\beta_1 = 0.9$, $\beta_2 = 0.999$. The learning rate is initially set to 10^{-4} and reduced to 10^{-6} after 20 epochs. The training generally converges after 40 epochs. The proposed network is implemented using TensorFlow② and trained with an NVIDIA GeForce GTX 1080TI GPU.

4 Experimental Results

In this section, we first give details about the testing sets, then demonstrate the effectiveness of the proposed modules via performing ablation study, and compare the proposed method with state-of-the-art methods finally. All the experiments are conducted for 4x upsampling.

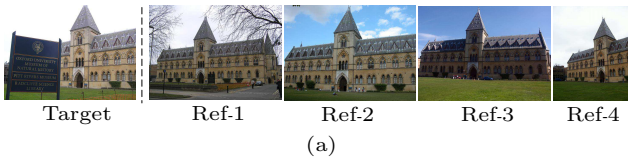
4.1 Testing Sets

Since our training set only contains landmark images, we evaluate the proposed method using three datasets to demonstrate the robustness of the trained model. The three datasets are Landmark10 from the work of [10], CUFED5 (containing 126 images) from the work of [11], and the Face50 dataset randomly selected from the VGGFace2 dataset [36]. The Landmark10 dataset contains 10 landmark images and its references are the same building captured by different people from various viewpoints at different time points. The CUFED5 dataset is collected from albums, which describes the most common events in our daily life, and the reference images describe the same event. Face50 is a face database randomly selected from the VGGFace2 dataset, which contains face images with large variations in pose, age, illumination, ethnicity, and profession. The three datasets represent three typical scenarios for RefSR. Some examples of our test images from the three test sets are presented in Fig.2. Fig.3 presents the reference images for test images from the three datasets. It can be observed that our test scenes have a large variance, and the reference images also vary in contents and viewpoints.

(a)

(b)

Fig.2. Some examples of testing images from (a) Landmark10 [10] and (b) CUFED5 [11] datasets.



(a)

(b)

(c)

Fig. 3. Reference images for one target image from (a) Landmark10 [10], (b) CUFED5 [11], and (c) Face50 [36] datasets.

4.2 Ablation Studies

In this subsection, we perform ablation study to demonstrate the effectiveness of the proposed CAM module and the progressively upsampling strategy.

②https://www.tensorflow.org/, Apr. 2020.

4.2.1 Ablation for CAM

To demonstrate the effectiveness of the proposed CAM, we compare it with the scheme by replacing the proposed CAM with SE [29] attention or removing CAM, i.e., without channel attention. Except for the attention module, all the other settings for the two variants are the same as those of our PCANet. Table 1 presents the ablation results for the proposed CAM module in terms of PSNR and SSIM [37] values on the three datasets. Note that the best results are highlighted in bold. All the values are calculated in the luminance channel. It can be observed that, the results of the scheme without channel attention (w/o CA, i.e., the baseline method) are the worst. Compared with the baseline method, the scheme with the SE module improves the average PSNR result by 0.27 dB. In contrast, the proposed method, i.e., using CAM as the attention module, improves the baseline method by 0.42 dB on average. We also present the visual comparison results in terms of PSNR/SSIM for the three schemes in Fig.4. Our method recovers the most details compared with the two variants.

Table 1. Ablation Study for the CAM Module

Algorithm	Landmark [10]	CUFED5 [11]	Face50 [36]
w/o channel attention	25.59/0.773 1	25.21/0.775 8	31.46/0.844 8
SE [29]	25.95/0.791 5	25.46/0.783 9	31.66/0.851 2
CAM(ours)	**26.23/0.803 0**	**25.60/0.794 0**	**31.71/0.851 6**

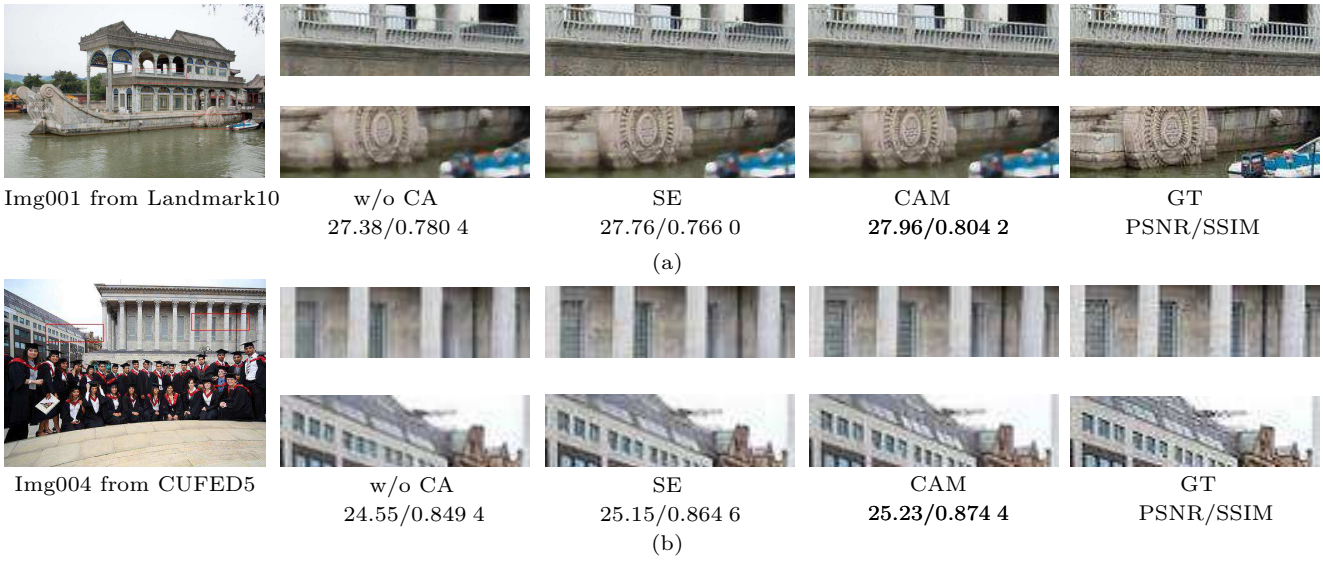


Fig.4. Visual comparison for ablation study of the CAM module. We show the close-up of the rectangle regions for better observation. (a) Img001 from Landmark10 and the visual comparison results. (b) Img004 from CUFED5 and the visual comparison results.

4.2.2 Ablation for Progressive Reconstruction

To demonstrate the effectiveness of the proposed progressive reconstruction strategy, we compare it with directly upsampling. Namely, we only utilize one level upsampling to upsample the LR input to the desired resolution. For a fair comparison, we set the network depth of the directly upsampling the same as our progressive upsampling to make the two networks have similar amount of parameters. Table 2 presents that the average PSNR result on the three datasets for directly upsampling is 27.54 dB, and our progressive upsampling (for 4x) strategy outperforms it by 0.3 dB. The visual comparison results are presented in Fig.5. It can be observed that the proposed progressive reconstruction strategy recovers dense structures while the one level upsampling can only recover limited structures since dense structures are hard to be predicted by one level upsampling.

Table 2. Ablation Study for Progressive Reconstruction

Algorithm	Landmark [10]	CUFED5 [11]	Face50 [36]
Bicubic	23.12/0.699 3	22.77/0.654 6	29.42/0.792 3
Directly upsampling	25.88/0.789 8	25.30/0.783 0	31.46/0.846 2
Progressively upsampling	**26.23/0.803 0**	**25.60/0.794 0**	**31.71/0.851 6**

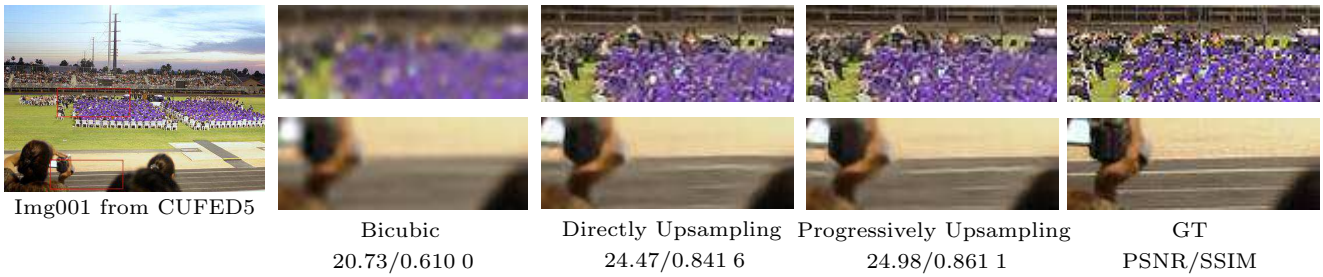


Fig.5. Visual comparison for ablation study of the progressively upsampling strategy. We show the close-up of the rectangle regions for better observation.

4.2.3 *Ablation for PCANet*

To demonstrate the effectiveness of the proposed PCANet, we compare it with directly merging the most similar HR patches for each LR patch together via averaging (denoted by patch averaging). Fig.6 presents the visual comparison results. It can be observed that the result of patch averaging contains many artifacts. In contrast, the result of proposed PCANet is sharp and clean. The reason is that the matched HR patches may be not similar to the LR input, as shown in Fig.6(c). Table 3 presents the quantitative comparison results. The proposed PCANet greatly outperforms the patch averaging strategy. This further demonstrates that the proposed PCANet can deal with HR patches with different similarity levels to the LR input. In summary, both objective and subjective comparisons prove that the proposed method is good at using input information and obtaining the best results.

4.3 Comparison with State-of-the-Art Methods

We compare the proposed PCAnet with the state-of-the-art SISR method EDSR [6] and RefSR methods, including one traditional patch matching and blending method Landmark [10] and three CNN-based methods, i.e., SRNTT [11], CrossNet [12] and IENet [16]. All the comparison results are obtained by the authors' codes. Since the original CrossNet model is trained using light-field images with small displacements, for a fair comparison, we retrain the CrossNet model using the aligned reference image. We choose the most similar image as its reference since it can only utilize one reference image. For SRNTT [11] and IENet [16], we directly utilizes their pretrained model for testing. For SRNTT, we present two kinds of results, i.e., SRNTT trained with perceptual and adversarial losses, and SRNTT-l_2 trained with l_2 loss which tends to have high PSNR results but smooth details. Therefore, for the visual comparison, we only present the first version SRNTT result.

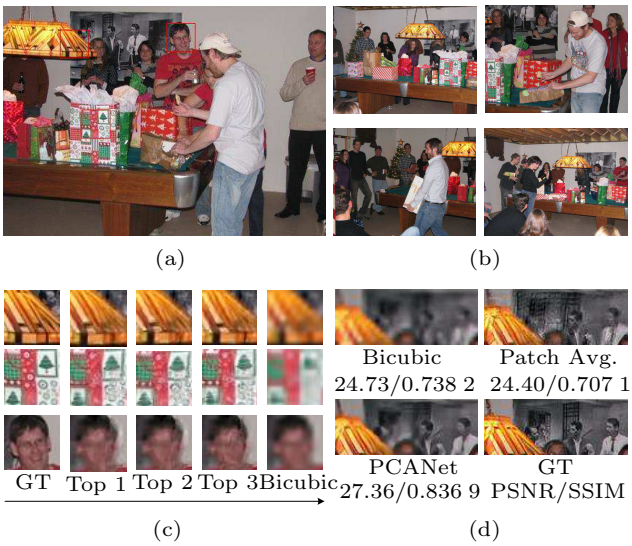


Fig.6. Visual comparison with the patch averaging result. (a) Ground truth, i.e., Img125 from CUFED5. (b) References of Img125. (c) Top three similar HR patches for three LR patches. (d) SR results generated by different methods.

Table 3. Quantitative Comparison Between Patch Averaging and the Proposed PCANet

Algorithm	Landmark [10]	CUFED5 [11]	Face50 [36]
Bicubic	23.12/0.699 3	22.77/0.654 6	29.42/0.792 3
Patch Averaging	24.76/0.720 8	23.70/0.710 4	30.13/0.811 9
PCANet	**26.23/0.803 0**	**25.60/0.794 0**	**31.71/0.851 6**

Table 4 presents the quantitative comparison results③ in terms of average PSNR and SSIM values on three datasets, i.e., Landmark10 [10], CUFED5 [11], and Face50 [36]. The best results are highlighted in

③The PSNR and SSIM results are different from those presented in SRNTT [11], since their PSNR and SSIM values are calculated in clipped luminance channel, whose values range from 15 to 235. If we calculate PSNR/SSIM in that way, the PSNR/SSIM values of SRNTT-l_2 and PCANet will be 26.00/0.765 1, and 26.90/0.808 4, respectively.

bold. It can be observed that the proposed PCANet outperforms the SISR method EDSR [6] by more than 1 dB on the Landmark10 dataset [10]. Compared with the second best method IENet [16], our method achieves more than 0.3 dB gain on all the three datasets.

Table 4. Comparison of Average SR Results on Three Datasets in Terms of Average PSNR/SSIM Values

Algorithm	Landmark [10]	CUFED5 [11]	Face50 [36]
Bicubic	23.12/0.639 3	22.77/0.654 6	29.42/0.792 3
EDSR [6]	24.93/0.738 8	25.04/0.759 3	31.34/0.836 0
Landmark [10]	24.76/0.720 8	23.70/0.710 4	30.13/0.811 9
CrossNet [12]	24.83/0.737 0	24.58/0.747 8	30.70/0.825 4
SRNTT [11]	23.73/0.687 5	24.05/0.708 5	29.62/0.799 1
SRNTT-l_2 [11]	24.58/0.723 3	24.57/0.742 8	30.78/0.833 7
IENet [16]	25.74/0.785 8	25.15/0.778 2	31.39/0.845 7
PCANet	**26.23/0.803 0**	**25.60/0.794 0**	**31.71/0.851 6**

Figs.7–9 present the visual comparison results for several test images on the three datasets, respectively. More visual comparison results are listed in the supplementary material④. It can be observed that the proposed PCANet recovers the most details, such as the building structures presented in Fig.7. In contrast, the result of EDSR has the least details since it only utilizes the information of the LR image. The result of CrossNet is a bit smooth since it cannot utilize the reference image with large displacements well. The

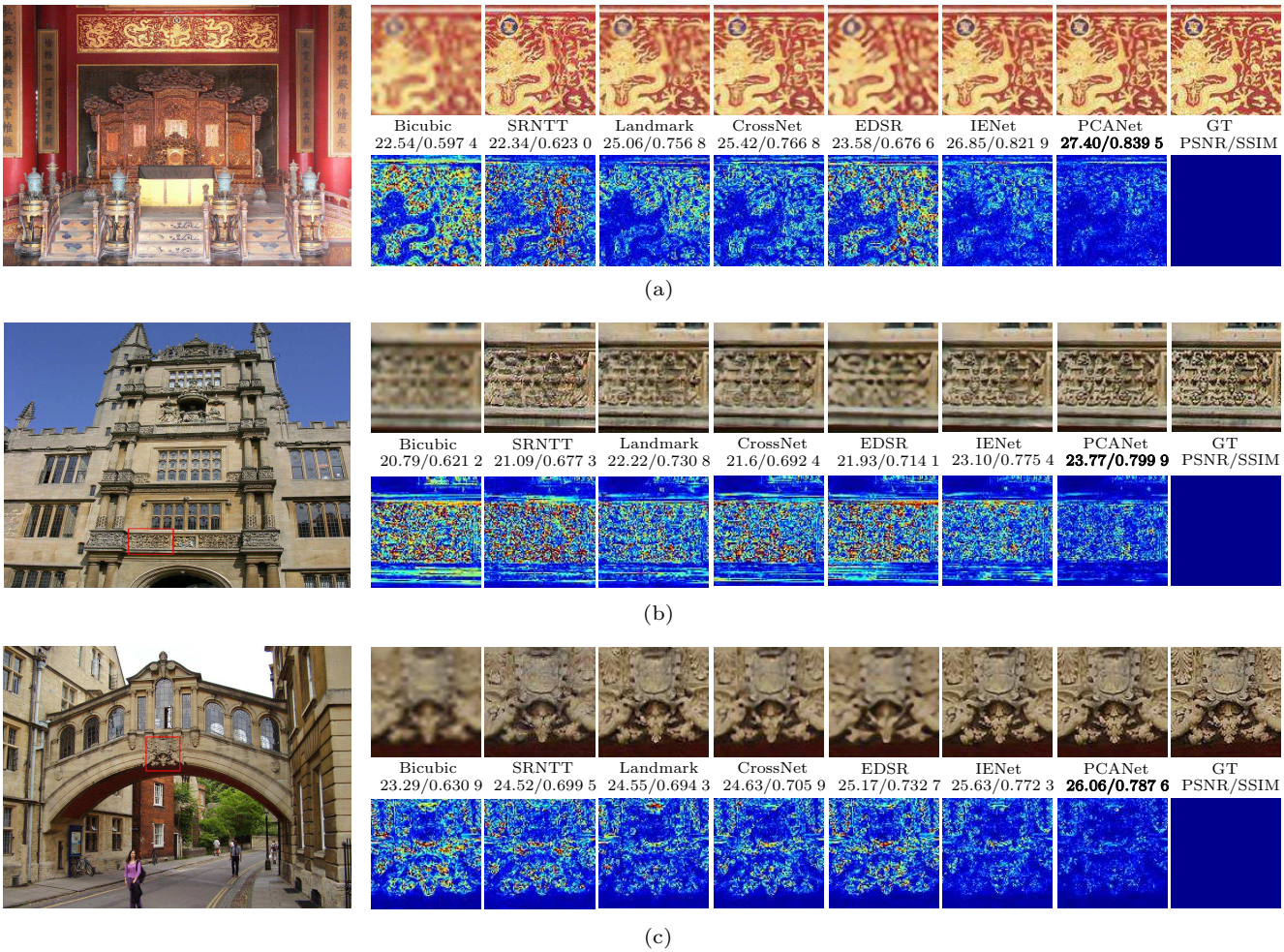


Fig.7. Comparison with competing SR methods on the Landmark10 [10] dataset in terms of PSNR/SSIM. For better observation, in each group, we present one highlighted region for each result (the top row) and provide the error heatmaps between the SR result and the ground truth (the bottom row). The hotter the color, the larger the errors. (a) Img002 from Landmark10 and the results. (b) Img008 from Landmark10 and the results. (c) Img009 from Landmark10 and the results.

④The supplementary material is avaliable at https://drive.google.com/open?id=1bVYl-rE5XDD182kP5YIb-9ECCLc5VdV8, Apr. 2020.

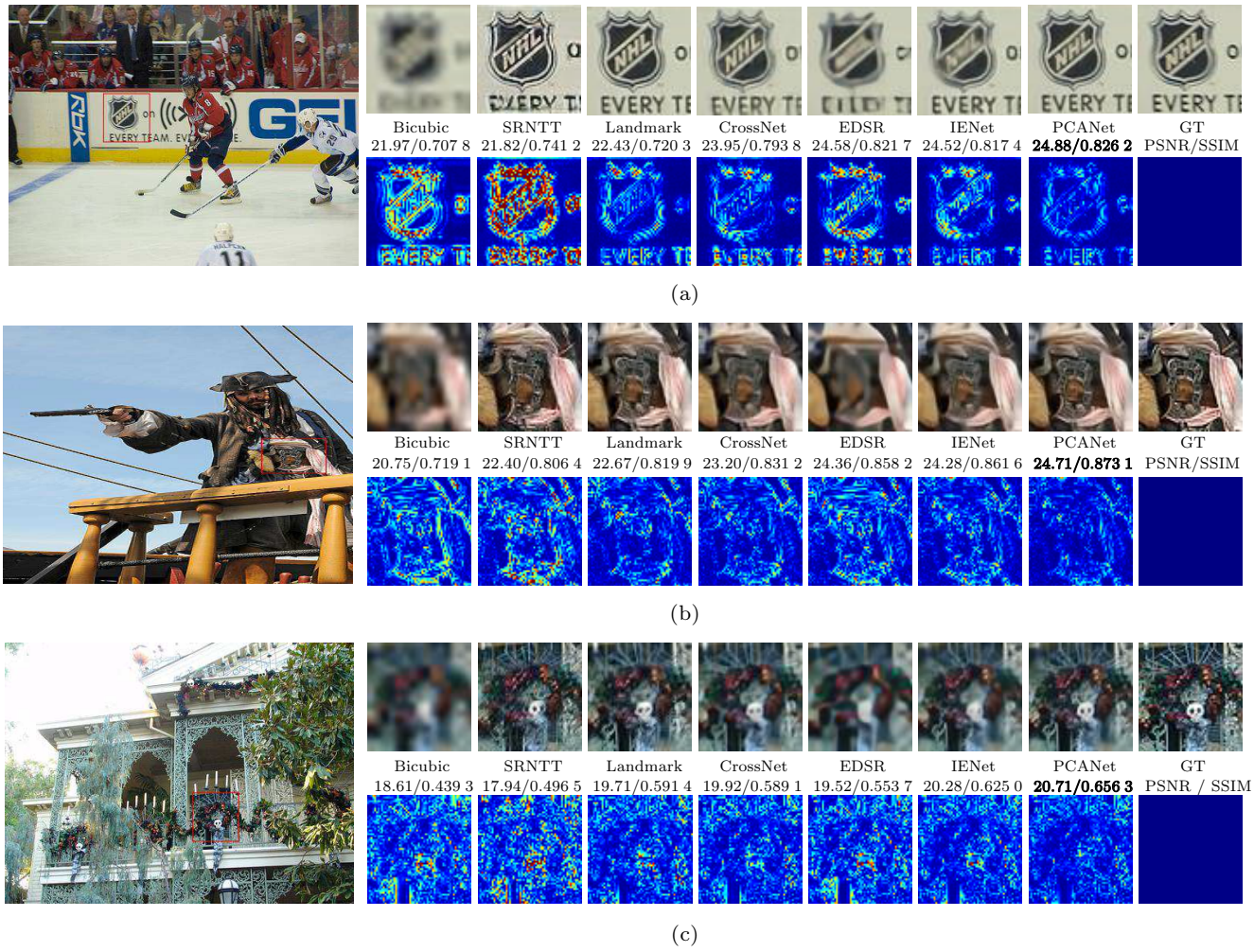


Fig.8. Comparison with competing SR methods on the CUFED5 [11] dataset in terms of PSNR/SSIM. For better observation, in each group, we present one highlighted region for each result (the top row) and provide the error heatmaps between the SR result and the ground truth (the bottom row). The hotter the color, the larger the errors. (a) Img002 from CUFED5 and the results. (b) Img006 from CUFED5 and the results. (c) Img007 from CUFED5 and the results.

result of SRNTT has rich high frequency details since it utilizes perceptual and adversarial losses during training. However, the recovered details deviate from the ground truth, which makes it have lower PSNR values. IENet generates good results for most images, but it also tends to blur the details for some images, such as the result presented in Fig.8 and Fig.9. In a word, the proposed method generates the best results in terms of both objective and subjective comparisons.

5 Conclusions

In this paper, we proposed a progressive channel attention network, PCANet, for the reference image guided super-resolution (RefSR). To fully explore the correlations between the reference patches and the LR input, we proposed a novel channel attention module, CAM, to assign higher weights to more correlated features. Different from the traditional SE module for channel attention, the proposed CAM estimates the channel weights via weightedly averaging the spatial features instead of using global averaging. We also proposed a progressive upsampling strategy for RefSR, which takes advantage of reference images in multiple scales. The proposed method outperforms state-of-the-art SR methods on three RefSR datasets in both subjective and objective measurements.

In the future, we would like to extend the channel attention module to more reference image guided processing applications, such as RefDenoise and RefDeblur. In addition, we would like to explore the application of RefSR in other kinds of images, such as SAR images, medical images and so on.

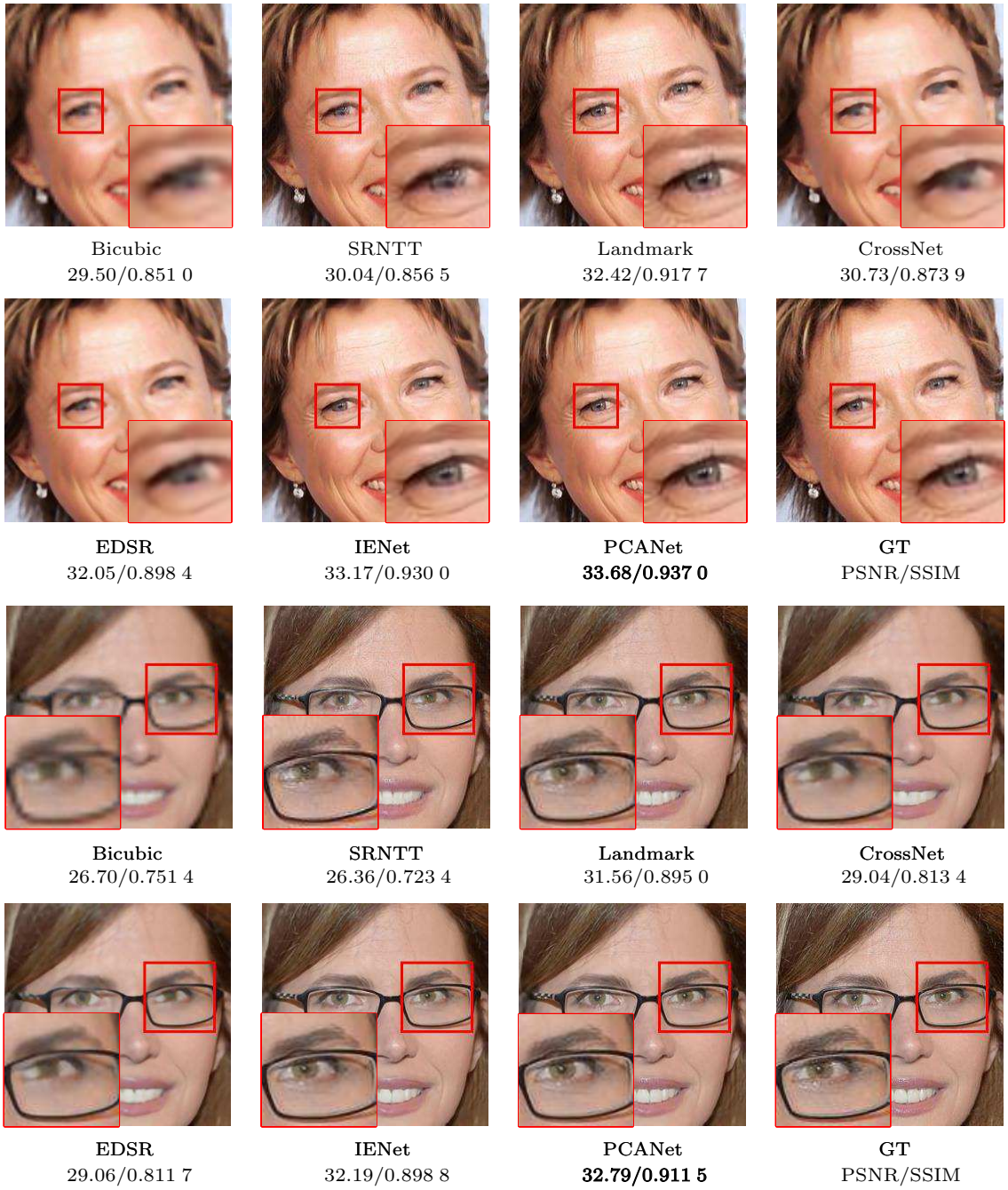


Fig.9. Comparison with competing SR methods on the Face50 [36] dataset in terms of PSNR/SSIM.

References

[1] Dong C, Loy C C, He K *et al.* Learning a deep convolutional network for image super-resolution. In *Proc. the 13th European Conference on Computer Vision*, September 2014, pp.184-199.

[2] Wang Z, Liu D, Yang J *et al.* Deep networks for image super-resolution with sparse prior. In *Proc. the 2015 IEEE International Conference on Computer Vision*, December 2015, pp.370-378.

[3] Kim J, Lee J K, Lee K M. Accurate image super-resolution using very deep convolutional networks. In *Proc. the 2016 IEEE Conference on Computer Vision and Pattern Recognition*, June 2016, pp.1646-1654.

[4] Lai W S, Huang J B, Ahuja N *et al.* Deep Laplacian pyramid networks for fast and accurate super-resolution. In *Proc. the 2017 IEEE Conference on Computer Vision and Pattern Recognition*, July 2017, pp.5835-5843.

[5] Tong T, Li G, Liu X *et al.* Image super-resolution using dense skip connections. In *Proc. the IEEE International*

Conference on Computer Vision, October 2017, pp.4809-4817.

[6] Lim B, Son S, Kim H *et al*. Enhanced deep residual networks for single image super-resolution. In *Proc. the 2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops*, July 2017, pp.1132-1140.

[7] Hu Y, Li J, Huang Y *et al*. Channel-wise and spatial feature modulation network for single image super-resolution. *IEEE Transactions on Circuits and Systems for Video Technology*. doi:10.1109/TCSVT.2019.2915238.

[8] Zhang Y, Li K, Li K *et al*. Image super-resolution using very deep residual channel attention networks. In *Proc. the 15th European Conference on Computer Vision*, September 2018, pp.294-310.

[9] Liu S, Gang R, Li C *et al*. Adaptive deep residual network for single image super-resolution. *Computational Visual Media*, 2019, 5(4): 391-401.

[10] Yue H, Sun X, Yang J *et al*. Landmark image super-resolution by retrieving web images. *IEEE Transactions on Image Processing*, 2013, 22(12): 4865-4878.

[11] Zhang Z, Wang Z, Lin Z *et al*. Image super-resolution by neural texture transfer. In *Proc. the 2019 IEEE Conference on Computer Vision and Pattern Recognition*, June 2019, pp.7982-7991.

[12] Zheng H, Ji M, Wang H *et al*. CrossNet: An end-to-end reference-based super resolution network using cross-scale warping. In *Proc. the 15th European Conference on Computer Vision*, September 2018, pp.87-104.

[13] Li Y, Dong W, Shi G *et al*. Learning parametric distributions for image super-resolution: Where patch matching meets sparse coding. In *Proc. the 2015 IEEE International Conference on Computer Vision*, December 2015, pp.450-458.

[14] Liu J, Yang W, Zhang X *et al*. Retrieval compensated group structured sparsity for image super-resolution. *IEEE Transactions on Multimedia*, 2017, 19(2): 302-316.

[15] Yang W, Xia S, Liu J *et al*. Reference-guided deep super-resolution via manifold localized external compensation. *IEEE Transactions on Circuits and Systems for Video Technology*, 2019, 29(5): 1270-1283.

[16] Yue H, Liu J, Yang J *et al*. IENet: Internal and external patch matching ConvNet for web image guided denoising. *IEEE Transactions on Circuits and Systems for Video Technology*. doi: 10.1109/TCSVT.2019.2930305.

[17] Zhang J, Gai D, Zhang X *et al*. Multi-example feature-constrained back-projection method for image super-resolution. *Computational Visual Media*, 2017, 3(1): 73-82.

[18] Zhao X, Wu Y, Tian J *et al*. Single image super-resolution via blind blurring estimation and anchored space mapping. *Computational Visual Media*, 2016, 2(1): 71-85.

[19] Glasner D, Bagon S, Irani M. Super-resolution from a single image. In *Proc. the 12th IEEE International Conference on Computer Vision*, September 2009, pp.349-356.

[20] Freeman W T, Jones T R, Pasztor E C. Example-based super-resolution. *IEEE Computer Graphics and Applications*, 2002, 22(2): 56-65.

[21] Yang J, Wright J, Huang T *et al*. Image super-resolution as sparse representation of raw image patches. In *Proc. the 2008 IEEE Conference on Computer Vision and Pattern Recognition*, June 2008.

[22] Yang J, Wright J, Huang T S *et al*. Image super-resolution via sparse representation. *IEEE Transactions on Image Processing*, 2010, 19(11): 2861-2873.

[23] Kim J, Lee J K, Lee K M. Deeply-recursive convolutional network for image super-resolution. In *Proc. the 2016 IEEE Conference on Computer Vision and Pattern Recognition*, June 2016, pp.1637-1645.

[24] Shi W, Caballero J, Huszár F *et al*. Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network. In *Proc. the 2016 IEEE Conference on Computer Vision and Pattern Recognition*, June 2016, pp.1874-1883.

[25] Ledig C, Theis L, Huszár F *et al*. Photo-realistic single image super-resolution using a generative adversarial network. In *Proc. the 2017 IEEE Conference on Computer Vision and Pattern Recognition*, July 2017, pp.105-114.

[26] He K, Zhang X, Ren S *et al*. Deep residual learning for image recognition. In *Proc. the 2016 IEEE Conference on Computer Vision and Pattern Recognition*, June 2016, pp.770-778.

[27] Goodfellow I, Pouget-Abadie J, Mirza M *et al*. Generative adversarial nets. In *Proc. the 2014 Annual Conference on Neural Information Processing Systems*, December 2014, pp.2672-2680.

[28] Johnson J, Alahi A, Li F F. Perceptual losses for real-time style transfer and super-resolution. In *Proc. the 14th European Conference on Computer Vision*, October 2016, pp.694-711.

[29] Hu J, Shen L, Sun G. Squeeze-and-excitation networks. In *Proc. the 2018 IEEE Conference on Computer Vision and Pattern Recognition*, June 2018, pp.7132-7141.

[30] Mansimov E, Parisotto E, Ba J L *et al*. Generating images from captions with attention. arXiv:1511.02793, 2015. https://arxiv.org/abs/1511.02793, Jan. 2020.

[31] Kim J H, Choi J H, Cheon M *et al*. Ram: Residual attention module for single image super-resolution. arXiv:1811.12043, 2018. https://arxiv.org/pdf/1811.12043.pdf, Jan. 2020.

[32] Woo S, Park J, Lee J Y *et al*. CBAM: Convolutional block attention module. In *Proc. the 15th European Conference on Computer Vision*, September 2018, pp.3-19.

[33] Lowe D G. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 2004, 60(2): 91-110.

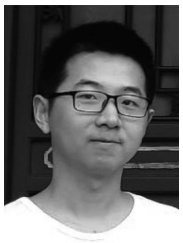
[34] Fischler M A, Bolles R C. Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, 1981, 24(6): 381-395.

[35] Kingma D P, Ba J. ADAM: A method for stochastic optimization. arXiv:1412.6980, 2014. https://arxiv.org/pdf/1412.6980.pdf, Jan. 2020.

[36] Cao Q, Shen L, Xie W *et al*. VGGFace2: A dataset for recognising faces across pose and age. In *Proc. the 13th IEEE International Conference on Automatic Face & Gesture Recognition*, May 2018, pp.67-74.

[37] Wang Z, Bovik A C, Sheikh H R *et al*. Image quality assessment: From error visibility to structural similarity. *IEEE Transactions on Image Processing*, 2004, 13(4): 600-612.

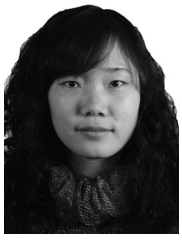
Huan-Jing Yue received her B.S. and Ph.D. degrees in electronic and information engineering from Tianjin University, Tianjin, in 2010 and 2015, respectively. She was an intern with Microsoft Research Asia, Beijing, from 2011 to 2012, and from 2013 to 2015. She visited the Video Processing Laboratory, University of California at San Diego, from 2016 to 2017. She is currently an associate professor with the School of Electrical and Information Engineering, Tianjin University, Tianjin. Her current research interests include image processing and computer vision. She received the Microsoft Research Asia Fellowship Honor in 2013 and was selected into the Elite Scholar Program of Tianjin University in 2017.



Sheng Shen received his B.S. degree in information engineering from Hefei University of Technology, Hefei, in 2018. He is currently pursuing his M.E. degree in electronic and information engineering at Tianjin University, Tianjin. His current research interests include image and video super resolution.

Jing-Yu Yang received his B.E. degree in automation from Beijing University of Posts and Telecommunications, Beijing, in 2003, and Ph.D. (Hons.) degree in automation from Tsinghua University, Beijing, in 2009. He has been a faculty member with Tianjin University, Tianjin, since 2009, where he is currently a professor with the School of Electrical and Information Engineering. He was with Microsoft Research Asia (MSRA), Beijing, in 2011, and the Signal Processing Laboratory, EPFL, Lausanne, Switzerland, in 2012, and from 2014 to 2015. His research interests include image/video processing, 3D imaging, and computer vision. He has authored or co-authored over 90 high quality research papers (including dozens of IEEE Transactions and top conference papers). As a co-author, he got the Best 10% Paper Award in IEEE VCIP 2016 and the Platinum Best Paper Award in IEEE ICME 2017. He served as special session chair in VCIP 2016 and area chair in ICIP 2017. He was selected into the program for New Century Excellent Talents in University (NCET) from the Ministry of Education of China, in 2011, the Reserved Peiyang Scholar Program of Tianjin University in 2014.

Hao-Feng Hu received his B.S. and Ph.D. degrees in electronic and information engineering from Nankai University, Tianjin, in 2002 and 2011, respectively. He visited the Institute of Optics, French National Center for Scientific Research, Paris, from 2011 to 2013. He is currently an associate professor with the School of Precision Instrument and Opto-Electronics Engineering, Tianjin University, Tianjin. His research interests include optical imaging and polarization imaging technologies.



Yan-Fang Chen is a Ph.D. student in the School of Electronic and Information Engineering, Tianjin University, Tianjin. Her current research interests include image processing and target recognition.