Neurocomputing xxx (xxxx) xxx

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

RBPNET: An asymptotic Residual Back-Projection Network for super-resolution of very low-resolution face image

Xiaozhen Chen[a], Xuebo Wang[a], Yao Lu[a,*], Weiqi Li[a], Zijian Wang[a,b], Zhuowei Huang[b]

[a] Beijing Laboratory of Intelligent Information Technology, Beijing Institute of Technology, Beijing, China
[b] China Central Television, Beijing, China

ARTICLE INFO

Article history:
Received 17 April 2019
Revised 17 August 2019
Accepted 27 September 2019
Available online xxx

Communicated by Prof. Liu Guangcan

Keywords:
Super-resolution
Very low-resolution face image
Residual learning
Back projection
Self-supervision

ABSTRACT

The super-resolution of a very low-resolution face image is a challenge task in single image super-resolution. Most of deep learning methods learn a non-linear mapping of input-to-target space by one-step upsampling. These methods are difficult to reconstruct a high-resolution face image from single very low-resolution face image. In this paper, we propose an asymptotic Residual Back-Projection Network (RBPNet) to gradually learn residual between the reconstructed face image and the ground truth by multi-step residual learning. Firstly, the reconstructed high-resolution feature map is projected to the original low-resolution feature space to generate low-resolution feature map (the projected low-resolution feature map). Secondly, the projected low-resolution feature map is subtracted by original feature map to generate low-resolution residual feature map. And finally, the low-resolution residual feature map is mapped to high-resolution feature space. The network will get a more accurate high-resolution image by iterative residual learning. Meanwhile, we explicitly reconstruct the edge map of face image and embed it into the reconstruction of high-resolution face image to reduce distortion of super-resolution results. Extensive experiments demonstrate the effectiveness and advantages of our proposed RBPNet qualitatively and quantitatively.

1. Introduction

In order to meet the demand of face detection and recognition, face image super-resolution (face SR) has been paid much more attention in recent years. It aims to reconstruct a high-resolution face image from a low-resolution face image and improve the visual effect and the accuracy of face detection and recognition.

Recently, learning-based face SR approaches [1–5] have shown better performance than traditional methods. However most current face SR methods can get pleasing results only when upscale factors is not large ($2\times$, $3\times$, $4\times$) and input low-resolution image has middle-level resolution (64×64, 32×32). When the resolution of an input face image is lower (16×16), the results of predicting high-resolution face images by these methods of one-step upsampling are unsatisfactory. As shown in Fig. 1(d), the reconstructed high-resolution face image by the state-of-the-art face super-resolution method [1] is visually inconsistent with the real one. Lai et al. [6] proposes deep laplacian pyramid network for fast and accurate super-resolution (LapSRN), which is used to reconstruct generic images, to progressively reconstruct the sub-band residuals of high-resolution images using more supervision information at each level in pyramid. However, LapSRN needs more ground truth as image labels and manipulates the prediction of the neural network, which limits the predictive power of neural networks. As shown in Fig. 1(c), retraining LapSRN using face dataset will generate a blurred face image.

We propose a novel super-resolution neural network RBPNet: an asymptotic Residual Back-Projection Network to super-resolve very low-resolution face image by gradually residual learning using self-supervision information. In order to avoid to introduce more supervising labels in the neural network, our intuition is using the original low-resolution face image as a reference. Inspired by the traditional super-resolution method of iterative back-projection [7], we project the reconstructed high-resolution feature map back to the original low-resolution feature space and then it is subtracted by the original low-resolution feature map to obtain a low-resolution residual feature map. Finally, the low-resolution residual feature map is mapped to high-resolution feature space. It is not enough to approximate the real high-resolution face image with only one-step residual learning, so we iteratively learn the residuals between the generated image and ground truth by multiple steps. Our network successfully super-resolve a very low-resolution

* Corresponding author.
E-mail address: vis_yl@bit.edu.cn (Y. Lu).

https://doi.org/10.1016/j.neucom.2019.09.079
0925-2312/

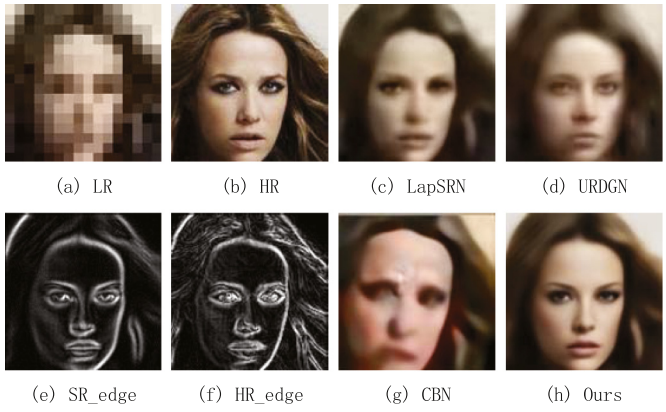

Fig. 1. Comparison with the state-of-the-art CNN based face super-resolution methods. (a) 16×16 LR input image. (b) 128×128 HR original image. (c) Result of deep Laplacian pyramid networks for fast and accurate Super-Resolution (LapSRN) [6]. (d) Result of Ultra-Resolving face images by Discriminative Generative Networks (URDGN) [1]. (e) The reconstructed high-resolution edge map by our RDBPNet. (f) Original high-resolution edge map. (g) Result of deep Cascaded Bi-Network for face hallucination (CBN) [2]. (h) Our result.

face image of 16×16 pixels to its $8\times$ larger version, as shown in Fig. 1(h).

Meanwhile, we take the global face structure into account instead of considering only local information to keep the reconstructed face image structure undistorted. Zhu et al. [2] propose deep Cascaded Bi-Network (CBN) to localize facial components in LR face images and then upsample them progressively. However, localizing these facial components with high accuracy is generally a difficult task in very low-resolution face images. As shown in Fig. 1(d), inaccurate prediction of facial components directly leads to distortion of reconstructed face images. Instead of localizing facial components, we use face edge map as global face structure information (see Fig. 1(f)). Edge prior has been used in image super-resolution to make the reconstructed high-resolution image sharper [8,9]. In contrast to previous methods, the edge map in our method not only can make the image sharper but also can be served as the global information to make the resultant image not distorted. We keep the reconstructed face image undistorted by explicitly predicting the edge map of the face image and embedding it into the reconstruction process of the face image. Besides, we also employ a discriminator network to differentiate the super-resolved images from original face images based on the Generative Adversarial Network (GAN) [10] structure.

Contributions of our work can be summarized as:

1) We present a novel end-to-end framework RBPNet to super-resolve very low-resolution face image. The network gradually learns high-resolution residual images by self-supervision mechanism, and finally approximate real high-resolution images.
2) The proposed network explicitly learns the edge map of high-resolution face image and embed it into the reconstruction of high-resolution face image to maintain the global structure of face image. We demonstrate the validity of the method qualitatively and quantitatively.

2. Related work

2.1. Single image super-resolution (SISR)

With the improvement of computing performance, many neural network based super-resolution methods have emerged and demonstrated better performance than traditional methods. Image super-resolution using deep convolutional networks (SRCNN) [11] is the pioneering work of deep learning for super-resolution reconstruction. It upsamples the low-resolution image to the desired size by bicubic interpolation, and then learn the mapping of low-resolution and high-resolution image patches through only three convolution layers. Real-time single image and video super-resolution using an Efficient Sub-Pixel Convolutional neural Network (ESPCN) [12] proposes a sub-pixel convolution layer to improve the computational efficiency of the network. Li et al. [13] present a two-channel convolutional neural network (SDSR) to restore the general outline of the image and detailed texture information simultaneously. Image super-resolution via a Densely Connected Recursive Network (DCRN) [14] is proposed to reconstruct high-quality images with fewer parameters and less computation time. So far, RCAN [15] (Very Deep Residual Channel Attention Networks) is performed best on single generic image super-resolution in PNSR and SSIM in case of $4\times$ super-resolution or less. Because they take interdependencies among channels into account, add channel attention mechanism to adaptively rescale channel-wise features. LapSRN [6] uses a cascading network to learn high-frequency residual details progressively, enabling $8\times$ super-resolution. EUSR [16] (Deep Residual Network with Enhanced Upscaling Module) proposes an enhanced upscaling module, which utilizes nonlinear operation and skip connection, performed well on $8\times$ super-resolution.

3. Proposed method

3.1. Back-projection

Back-projection [7] is well known as the efficient iterative procedure to minimize the reconstruction error. Recent works [17,18] have proven the effectiveness of back projection on super-resolution. Zhao et al. [17] propose a method to refine high-frequency texture details with an iterative projection process. Deep Back-Projection Networks for super-resolution (DBPN) [18] constructs mutually-connected up- and down-sampling to imitate the process of super-resolution and image degradation to learn more about degradation patterns.

Different from the above methods, we propose RBPNet to project the reconstructed high-resolution feature map back to the original low-resolution feature space to supervise reconstructed HR face images (self-supervision mechanism) and learn to map the residuals of low-resolution feature space to high-resolution feature space to complement the residual information onto the high-resolution feature map, which is a process from coarse to fine.

Our network structure mainly consists of generator and discriminator. The generator is asymptotic residual network with embedding edge map and is composed of two branches: (1) an upsamling branch that upsamples original LR face images and their LR edge maps simultaneously. (2) A edge map embedded branch that explicitly embeds edge map in the reconstruction process. The discriminator is similar to SRGAN [19]. The proposed network is shown in Fig. 2.

3.2. Asymptotic Residual Back-Projection Network

Let I_{LR} (the concatenation of *LR* face images and *LR* edge maps) and I_{SR} denote the input and output images respectively. In order to improve the expressive ability of residual learning, we map it to the high-dimensional feature space. As shown in Fig. 2(a), we extract shallow feature F_0^L from I_{LR}, so that:

$$F_0^L = f_{shallow}(I_{LR}) \tag{1}$$

Next, we upsample F_0^L to the desired size (128×128) using deconvolution. (More detailed deconvolution operations can be found in the experimental section.) In order to get the difference between

Please cite this article as: X. Chen, X. Wang and Y. Lu et al., RBPNET: An asymptotic Residual Back-Projection Network for super-resolution of very low-resolution face image, Neurocomputing, https://doi.org/10.1016/j.neucom.2019.09.079

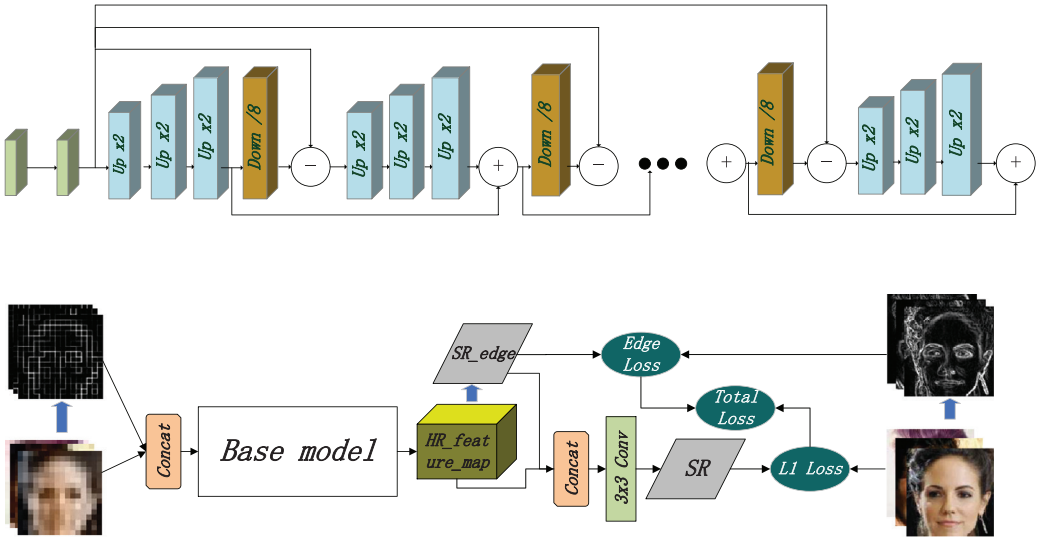

Fig. 2. The architecture of our network. (a) Base model: asymptotic Residual Back-Projection Network. (b) The network of embedding edge maps explicitly into the reconstruction of face images.

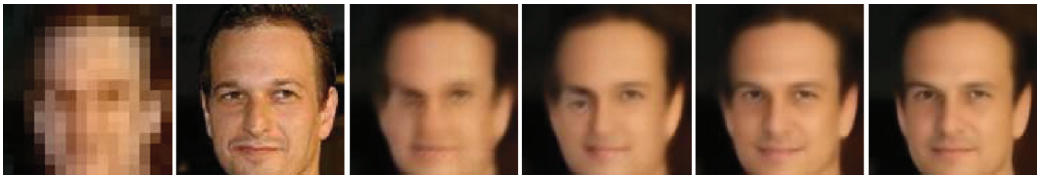

Fig. 3. Comparisons of different losses for the super-resolution by our RBPNet. (a) Original LR inputs. (b) Original HR images. (c) $L = L_p = L_2$ (d) $L = L_p = L_1$. (e) $L = L_p + L_e = L_1 + L_e$. (f) $L = L_p + \alpha L_f + \beta L_g + L_e$.

real HR feature map F_1^H and reconstructed HR feature map F_0^H, we use a simple downsampling convolution to project reconstructed HR feature map F_0^H back to the LR feature map F_1^L:

$$F_0^H = f_{up}(F_0^L; \alpha_0) \tag{2}$$

$$F_1^L = f_{down}(F_0^H; \theta_0) \tag{3}$$

where α_i and θ_i are parameters of convolution or deconvolution in i_{th} step. We will get the difference (also known as residual) of original F_0^L and F_1^L. Then we learn residual mapping to map the residual in LR feature space to HR feature space and we will get the difference e_1 between F_1^H and F_0^H. The final output is the summation of F_0^H and the difference e_1:

$$e_1 = f_{up}((F_1^L - F_0^L); \gamma_0) \tag{4}$$

$$F_1^H = F_0^H + e_1 \tag{5}$$

where γ_i denotes deconvolution parameters in i_{th} step. We only show one-step residual learning.

Asymptotic residual back-projection aims to learn the residual between the real high-resolution feature map and reconstructed high-resolution feature map, described as $F_i^H - F_{i-1}^H$. In the first step, the real residual is $F_1^H - F_0^H$. F_1^H is the real high-resolution feature map which is the result we desire, but it cannot be obtained directly. On the contrary, the real low-resolution feature map can be obtained. So we project the learned high-resolution feature map F_0^H back to low-resolution space, and the residual of low-resolution feature map is $F_1^L - F_0^L$. In the process of back-projection, it is inevitable to introduce error. So we need multi-step residual learning to reduce errors and make the reconstructed HR feature map approximate to the real HR feature map. Multi-step residual learning is to repeat the operations of Eqs. (2)–(5). Each residual reference value is always the original LR feature map F_0^H, as shown in Fig. 4(b). The final residual of our predicted $\hat{F}_n^H$ can be described as follow:

$$\hat{F}_n^H = F_{n-1}^H + f_{up}((F_n^L - F_0^L); \gamma_i) \tag{6}$$

where n denotes the number of steps of residual learning.

3.3. Edge map embedding network

In order to ensure structure consistency between LR face images and HR face images, we use edge map as prior. In contrast to the other priors such as textures that are usually difficult to recover after image degradation, edges are much easier to detect in LR images. As shown in Fig. 3(c)–(e), embedding edge map into the reconstruction of high-resolution face images effectively reduces distortion of super-resolution results. To extract edges, we first apply an off-the-shelf edge detector (Sobel in this paper) on LR face images and HR face images respectively. As shown in Fig. 2(b), we concatenate extracted LR edge maps with LR face images.

RBPNet reconstructs HR images and edge maps simultaneously. Reconstructing high-resolution feature maps includes two steps: (1) HR image recovery, (2) edge prediction in the HR image. We add an edge constraint by comparing the HR edge map extracted from HR feature maps with the original HR edge map extracted from original HR images.

$$F_{edge} = f_{edge}(F_{output}; \mu) \tag{7}$$

Where μ denotes the parameter of SR edge extracted layer. Finally, we reconstruct high-resolution images with SR edge map embedding:

$$I_{SR} = f_{reconstruction}([F_{edge}, F_{output}], z) \tag{8}$$

Where z denotes the parameter of reconstruction layer.

Please cite this article as: X. Chen, X. Wang and Y. Lu et al., RBPNET: An asymptotic Residual Back-Projection Network for super-resolution of very low-resolution face image, Neurocomputing, https://doi.org/10.1016/j.neucom.2019.09.079

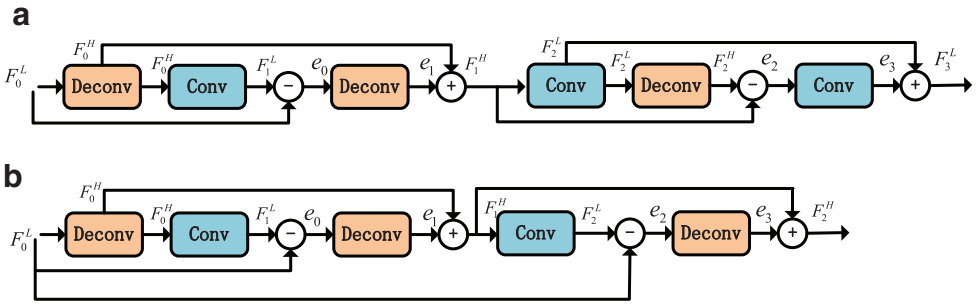

Fig. 4. difference of the network structure between DBPN and our RBPNet. (a) The up- and down-projection unit in DBPN [18]. (b) The residual back-projection unit in our RBPNet.

3.4. Discriminator network

Recent works [1,19,20] demonstrate that only using L_2 loss between the restored face images and the ground truth of HR face images usually fails to recover the sharp or high-frequency details. Therefore, we design a discriminator network to distinguish whether a upsampling face image is similar to authentic face image or not. As shown in Fig. 3(e) and (f), the discriminator network makes the hallucinated faces sharper and more realistic.

The discriminator network is composed of convolutional layers and fully connected layers which is same as [19]. Input the restored face images by our upsampling network and real face images into discriminator network separately to update the discriminative loss. In this method, we can reconstruct more realistic HR face images.

3.5. Difference between DBPN and RBPNet

Haris et al. [18] propose DBPN to super-resolve generic image and get a relatively good result. And it also performs well on very low-resolution face images, as shown in Table 2. Though DBPN and RBPNet all use the idea of back-projection [7], there are some differences between them:

On reconstruction strategy, DBPN constructs mutually-connected up- and downsampling to imitate the process of super-resolution and image degradation, thereby to learn nonlinear relation of LR and HR image. But we employ asymptotic residual learning to compensate the errors caused by back-projection in multiple steps, and the compensated residual will close to the residual of real high-resolution feature map.

For the details of back-projection, the error reference of each up (or down) back-projection unit is the output value of the last down (or up) back-projection unit in DBPN and transfer those projection errors to every stage by dense connection. For instance in Fig. 4(a), the reference value of F_2^H is F_1^H, which is reconstructed from last Up-Projection Unit. We think that F_1^H is difficult to be reconstructed accurately because of the difficulty of upsampling large multiples in one-step, which will introduce errors to the next projection unit. Unlike DBPN, we propose RBPNet to progressively learn residual using back-projection as shown in Fig 4(b). We use the low-resolution feature F_0^L as a reference and gradually approximate the real high-resolution image feature map F_n^H.

In addition, our network with 12.44M parameters and 8.28 GFLOPs (Floating Point Operations) is lighter than DBPN with 32.45M parameters and 184.15 GFLOPs. And we achieve better results than DBPN on the three test datasets (CelebA, Helen and AFLW).

3.6. Loss function

In this section, we describe the loss function used in our network. To illustrate simplicity, let x, y and $\hat{y}$ denote LR image, restored HR image and the ground truth image. Let x^e, y^e and $\hat{y}^e$ denote LR feature map, restored HR edge map and the ground truth of edge map.

Pixel-wise loss: In order to ensure consistency of image intensities between the estimated HR image and the ground truth, we employ a pixel-wise loss. Instead of using the Euclidean distance, also known as pixel-wise l_2 loss, which tends to output over-smoothed results, we apply pixel-wise l_1 loss to enforce the similarity of restored HR images and the ground truth, as shown in Fig. 3(c) and (d).

$$\mathcal{L}_p(y, \hat{y}; w_g) = \frac{1}{N}\sum_{i=1}^{N}|\hat{y}_i - y_i| = \frac{1}{N}\sum_{i=1}^{N}|G_w(x_i) - y_i| \tag{9}$$

where $\hat{y}_i$ and $G_w(x_i)$ both represent the restored face images by our generate network, w_g is the parameters of generate network, x_i and y_i denote the LR input face image and its HR ground-truth counterpart, respectively, N is the number of training samples in each batch.

Feature-wise loss: Only using pixel-wise loss function to handle the ill-posed inverse problem will lose high-frequency details such as texture. In order to generate visually superior image, we also use feature-wise loss functions based on Euclidean distances computed in the feature space as follow:

$$\mathcal{L}_f(y, \hat{y}; w_g) = \frac{1}{N}\sum_{i=1}^{N}||\psi(G_w(x_i)) - \psi(y_i)||_2^2 \tag{10}$$

where $\psi(.)$ denotes feature maps of a layer in VGG-16 [21].

Discriminator loss: When we input a low-resolution face image, there are still many solutions that satisfy the optimization of pixel-wise loss and feature-wise loss we mentioned above. Because the optimization-based super-resolution method itself is still an uncertain problem. In order to output more realistic face images in many possible face images satisfying minimize above all loss functions, we employ a discriminator network to distinguish whether a upsampling face images is similar to authentic face images. The goal of discriminator network is to be able to distinguish the upsampled face images from the ground truth. Therefore, we maximize the loss $\mathcal{L}_D$ as follow:

$$\mathcal{L}_{\mathcal{D}}(y, \hat{y}; w_d) = log\mathcal{D}_d(y_i) + log(1 - \mathcal{D}_d(\hat{y}_i)) \tag{11}$$

where w_d represents the parameters of the discriminator network $\mathcal{D}$.

Adversarial loss: In order to fool the discriminator network, our generation network should produce face images that are as close as possible to the real images. So the generate loss, also known as adversarial loss as follows:

$$\mathcal{L}_g(y, \hat{y}; w_g) = log(1 - \mathcal{D}_d(G_w(x_i))) \tag{12}$$

Edge-wise loss: We explore the face structure information during super-resolution to keep the restored face image structure undistorted. We embed edge maps to the process of super-resolution as

Please cite this article as: X. Chen, X. Wang and Y. Lu et al., RBPNET: An asymptotic Residual Back-Projection Network for super-resolution of very low-resolution face image, Neurocomputing, https://doi.org/10.1016/j.neucom.2019.09.079

Table 1
Network configuration.

	DBPN	RBPNet-B	RBPNet-L	RBPNet-F
Input (16×16 pixels)				[Input, edge map]
Extract features	Conv(3-1-1,256) Conv(1-1-0,64)	Conv(3-1-1,256) Conv(1-1-0,64)	Conv(3-1-1,256) Conv(1-1-0,64)	Conv(3-1-1,256) Conv(1-1-0,64)
Up($\times 7$)	Deconv(12-8-2,64) Conv(12-8-2,64) Deconv(12-8-2,64) Conv(12-8-2,64)	**Deconv(12-8-2,64)**	**Deconv(6-2-2,64)** **Deconv(6-2-2,64)** **Deconv(6-2-2,64)**	Deconv(6-2-2,64) Deconv(6-2-2,64) Deconv(6-2-2,64)
Down($\times 6$)	Deconv(12-8-2,64) Conv(12-8-2,64)	**Conv(12-8-2,64)**	Conv(12-8-2,64)	Conv(12-8-2,64)
Extract edge	✗	✗	✗	**Conv(3-1-1,1)**
Reconstruction	Conv(3-1-1,3)	Conv(3-1-1,3)	Conv(3-1-1,3)	Conv(3-1-1,3)
	Output (128×128 pixels)			[Output, edge map]

Table 2
Quantitative evaluation of state-of-the-art face super-resolution algorithms: average PSNR/SSIM/IFC. Bold text indicates the best performance and italic text indicates the second best performance.

	Methods	Biucbic	LapSRN [6]	DBPN [18]	URDGN [1]	CBN [2]	RCAN [15]	EUSR [16]	Ours
CelebA	PSNR	22.2025	23.9884	24.0100	23.6326	23.8004	*24.2301*	24.1106	**24.2391**
	SSIM	0.5653	0.6810	0.6812	0.6710	0.6723	*0.6918*	0.6886	**0.6921**
	IFC	0.4852	0.8965	0.9160	0.8122	0.8654	*0.9388*	0.9344	**0.9532**
Helen	PSNR	22.0944	23.0854	23.2108	22.9651	22.8954	23.6737	*23.7053*	**23.8402**
	SSIM	0.5984	0.6591	0.6621	0.6188	0.6150	0.6877	*0.6906*	**0.6932**
	IFC	0.7581	0.9329	0.9718	0.9015	0.8542	0.9778	*1.0286*	**1.2251**
AFLW	PSNR	21.7868	22.4925	22.5681	22.0664	21.9827	*22.7237*	22.7161	**22.8270**
	SSIM	0.5642	0.6108	0.6224	0.6102	0.5917	0.6360	*0.6389*	**0.6419**
	IFC	0.4597	0.7115	0.7261	0.7085	0.6984	*0.7342*	0.7158	**0.7377**

the global structure information in our network. We estimate the restored edge map with the real one extracted from ground truth as follows:

$$\mathcal{L}_e(y^e, \hat{y}^e; w_g) = \frac{1}{N}\sum_{i=1}^{N} ||H(\hat{y}_i) - H(G_w(x_i))||_2^2 \tag{13}$$

where H is Sobel operator in our experiment.

Training details: We use $\mathcal{L}$ as the objective function of our generated network to update w_g, including $\mathcal{L}_p, \mathcal{L}_f, \mathcal{L}_g, \mathcal{L}_e$:

$$\mathcal{L} = \mathcal{L}_p + \alpha\mathcal{L}_f + \beta\mathcal{L}_g + \mathcal{L}_e \tag{14}$$

where α, β are the trade-off weights. We set α and β to 0.01 experientially. In order to training our discriminator network $\mathcal{D}$ conveniently, we take the Eq. (11) to the opposite and use the gradient descent method to minimize it. Specifically, we use Adam optimization algorithm [22] to update the parameters w_g and w_d. The discriminator network and generation network are trained in an alternating fashion. The initial learning rate is set to 10^{-4} and decreased by a factor of 10 every 3×10^4 iterations.

4. Experiment result

4.1. Implementation details

We set the convolution kernel size to 3×3, padding to 1 and stride to 1 in the feature extraction layers (followed by a convolution layer with a kernel size of 1×1), high-resolution edge map extraction layer and reconstruction layer. We set all the convolution kernel size to 12×12, padding to 2 and stride to 8 in downsampling layer to downsample feature maps by a factor of eight. We use 6×6 convolution kernel size with 2 striding and 2 padding in pyramid upsampling layer. More detailed parameter settings are shown in Table 1. All convolution layers are followed by PReLu activation function [23] (no activation function in the reconstruction layer). We set the number of channels for all intermediate convolution layers to 64. We use the method described in [23] to initialize the weight of the convolution filters. The weights of the transposed convolution filters are initialized from bilinear interpolation kernel. We use batch size of 8 with pixels of 16×16 for LR image, while HR image size is 128×128 for training. All experiments were conducted using Pytorch on NVIDIA TITAN X GPUs.

4.2. Training and testing datasets

Our experiments are performed on three datasets: CelebA [24], Helen [25] and AFLW [26]. CelebA is a large-scale face dataset with about 0.2 million 128×128 face images. We randomly sample 5000 images as verification image set, 1000 images as test image set, and the other images as training image set. Helen dataset consists of 2000 training images and 330 test images with highly accurate, detailed, and consistent annotations of the primary facial components. According to the annotations, we crop the face image of 128×128 pixels from each image. Since the training set of Helen has few images, networks trained on 2000 training images will overfit. So we randomly sample 50,000 images from CelebA together with Helen's training set to train neural network. Specifically, we use 52,000 training images to train the neural network and 330 test images to test. AFLW has 13.23 thousands of face images from 1680 persons, which are annotated with up to 21 landmarks per image. We also crop the face image of 128×128 pixels and train the networks with randomly sampling 50,000 images from CelebA and 13,000 images from AFLW. Specifically, we use 63,000 training images to train the neural network and 230 test images to test. To produce LR images of 16×16 pixels, we use bicubic to downsample the HR images by a factor of eight.

We compare the proposed RBPNet with two state-of-the-art face super-resolution algorithms: CBN [2], URDGN [1], and four generic image super-resolution methods: EUSR [16], RCAN [15], DBPN [18], LapSRN [6]. We evaluate the reconstructed SR images with three commonly used image quality metrics: PSNR, SSIM and IFC. All of the compared methods are tested on the three test datesets (CelebA, Helen, AFLW). DBPN [18], LapSRN [6], EUSR, RCAN

Please cite this article as: X. Chen, X. Wang and Y. Lu et al., RBPNET: An asymptotic Residual Back-Projection Network for super-resolution of very low-resolution face image, Neurocomputing, https://doi.org/10.1016/j.neucom.2019.09.079

Table 3
Our method is compared with DBPN on PSNR, SSIM and parameters. All tests are performed on the CelebA testset.

	PSNR	SSIM	Parameters
DBPN	24.0100	0.6812	22.13M
RBPNet-B	24.1621	0.6897	16.97M
RBPNet-L	24.2305	0.6901	**11.32M**
RBPNet-F	**24.2421**	**0.6935**	12.44M
RBPNet-All	24.2391	0.6921	12.44M

were trained for super-resolving generic images, so we retrain them on the three training sets to suit better for face images.

4.3. Model analysis

Comparison with DBPN: In order to clearly explain the difference between our proposed model and DBPN, and demonstrate the validity of our model, we construct RBPNet-B as our base model. For a fair comparison, we removed the dense connection in DBPN. The specific network configuration is shown in Table 1. Here, let $(de)conv(k-s-p,c)$ be a (de) convolution layer, where k, s, p denotes (de)convolution kernel size, (de)convolution stride and padding size respectively, c denotes the number of (de)convolution kernel. Our RBPNet-B is composed of four parts like DBPN: extracting features, upsampling, downsampling, reconstruction. Note that the up- and down-sampling layers are alternated. In order to enlarge the original image, the number of upsampling layers are one more than the number of downsampling layers. Our RBPNet-B is the same as the DBPN, using 7 upsampling layers and 6 downsampling layers.

In order to make a fair comparison with DBPN, our network configuration is basically the same as DBPN. The difference is that we only use one (de)convolution layer in up- and down-sampling layers. Note that we did not introduce adversarial training here.

The main difference between RBPNet-B and DBPN is the *data flow*. As shown in Fig. 4, we use the original low-resolution feature map as the supervised information, continuously supervise the high-resolution feature map obtained by upsampling and complement the residual information onto the high-resolution feature map, which is a process from coarse to fine. However, DBPN constructs mutually-connected up- and down-sampling to imitate the process of super-resolution and image degradation to learn more about degradation patterns, as the author of DBPN mentioned in the abstract.

As shown in Table 3, our RBPNet-B has about 20% fewer parameters. Through the improvement of the data flow, our RBPNet-B gains 0.15 dB higher than DBPN on PSNR using fewer parameters. This evidence show that our network is more effective than DBPN to super-resolve a very low-resolution face image of 16×16 pixels to its $8\times$ larger version in a coarse-to-fine manner. As shown in Fig. 5, we visualize activation maps in upsampling layers. All activation maps are sampled from the 52nd channel of the upsampling layers, whose contrast is more strongly suitable for presentation. Upsampling activation maps at different stages show that high-resolution feature maps are generated from coarse to fine, which is in line with the original intention of design

Pyramid upsampling: In order to further reduce the difficulty of upsampling on the network, we improve the one-step upsampling in the upsampling layers to multi-step pyramid upsampling. Here, rather than using 12×12 deconvolution kernel, we use three 6×6 deconvolution kernel to magnify low-resolution face images by 8 times through three steps. The specific network parameters are shown in the Table 1, which we call as RBPNet-L. We incorporate three non-linear rectification layers instead of a single one, which makes the decision function more discriminative. In addition, we have further reduced the parameters of the network: assuming that both the input and the output of a three-layer 6×6 deconvolution stack has C channels ($C=64$ in our network), the stack is parameterized by $3\cdot6^2\cdot C^2=108C^2$ weights; at the same time, a single 12×12 deconvolution layer would require $12\cdot C^2=144C^2$ parameters. As shown in Table 3, pyramid upsampling achieves better results with fewer parameters than using one-step upsampling. Note that we also do not introduce adversarial training here.

Edge map embedding: We refer the network embedding the edge map as RBPNet-F. As shown in Table 1, RBPNet-F has four input channels, including three channels of input image and one low-resolution edge map. Before reconstructing the high-resolution image, we extract the edge map from the high-resolution feature maps, use the ground truth of edge map as the supervised information, and finally embed the high-resolution edge map into the reconstruction of the high-resolution image. In Fig. 3, it is shown that embedding the edge map explicitly into the reconstruction of the high-resolution image can reduce the distortion. Note that we also do not introduce adversarial training here.

Adversarial training: We incorporate a discriminative objective into our network to force super-resolved HR face images to lie on the manifold of real face images, which we call RBPNet-All. The discriminative network is same as SRGAN [27]. As it is not our contribution, we did not do more experiments about discriminative network. As shown in Fig. 3(f) and Table 3, although the effect of increasing adversarial training on PSNR is not obvious, the generated image is more realistic.

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

Table 2 shows quantitative comparisons for $8\times$ face super-resolution, where the best algorithm has been highlighted on Table 2. From Table 2, we can find that our RBPNet (RBPNet-All) achieves better performance than the existing methods in PSNR, SSIM and IFC. Our method significantly advances in IFC, with the improvement margin of 0.14, 0.09 compared with URDGN and CBN in CelebA dataset. Compared with the best method of face image super-resolution URDGN, our model is marginally higher by 0.02 and 0.43 dB in SSIM and PSNR respectively in CelebA dataset. When compared with the three methods (EUSR, DBPN, and LapSRN), which all aim to solve the problem of $8\times$ generic image super-resolution, we find that our method performs better than them for super-resolution of very low-resolution images in all three datasets. In addition to the methods described above, our method is comparable with RCAN, which achieves the state-of-the-art in super-resolution. Specifically, our method achieves 24.2391 dB for PSNR, 0.6921 for SSIM and 0.9532 for IFC, almost +0.01 dB, +0.001 and +0.02, respectively, slightly better than RCAN results in CelebA dataset.

Although EUSR and RCAN perform well on generic image, the super-resolution of very low-resolution face image need more supervised information to reconstruct high-resolution image by multi-step upsampling. For the reconstruction of very low-resolution image, it is difficult to keep content consistency and structural similarity. Our approach addresses the two problems in two ways. (1) The original low-resolution image feature map is taken as the supervision information of upsampling operation in each stage, to ensure that the high-resolution image feature map from each upsampling is gradually close to the real high-resolution image feature map. (2) In the process of high-resolution image reconstruction, image edge constraints are added to ensure the consistency between reconstructed high-resolution edge image and original edge image.

In Fig. 6, we show visual comparisons for $8\times$ face super-resolution on CelebA, Helen and AFLW test sets. The reconstructed face images by CBN [2] (see Fig. 6(h)) look incompatible compared

Please cite this article as: X. Chen, X. Wang and Y. Lu et al., RBPNET: An asymptotic Residual Back-Projection Network for super-resolution of very low-resolution face image, Neurocomputing, https://doi.org/10.1016/j.neucom.2019.09.079

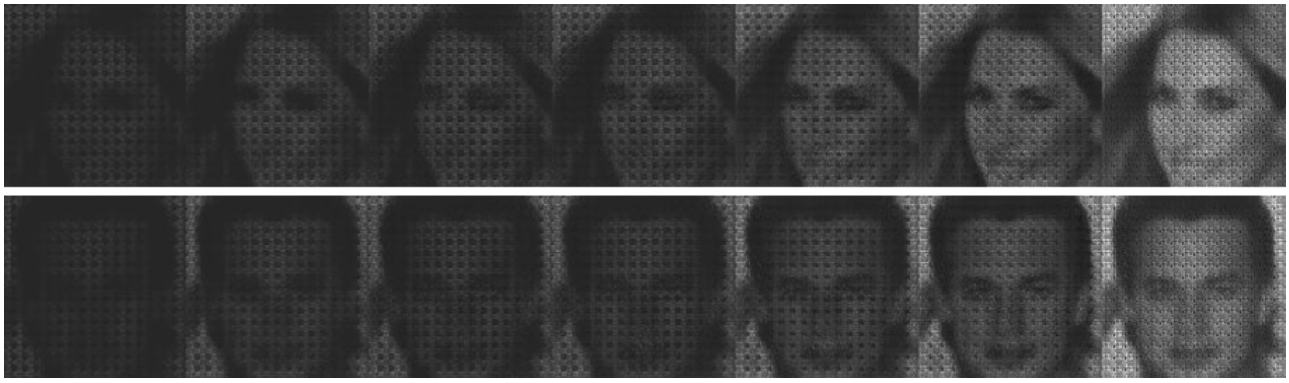

Fig. 5. Sample of activation maps from up-projection units in RBPNet-B where n=7. Each feature has been enhanced using the same grayscale colormap for visibility.

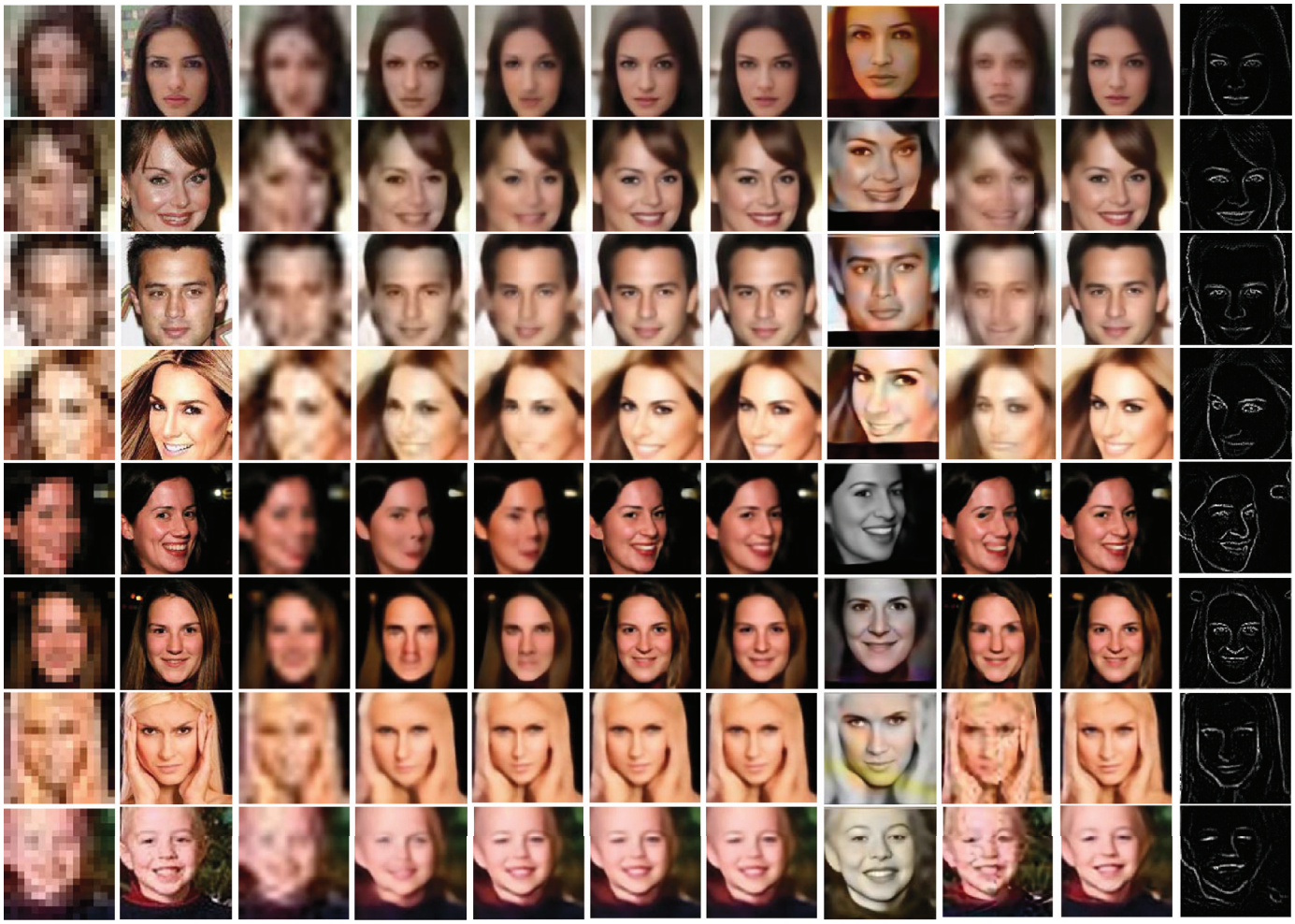

Fig. 6. Visual comparison for 8× face SR on CelebA, Helen and AFLW test sets. The first four lines are the test results on the CelebA test set. Lines 5 and 6 are the test results on Helen test set. The rest two lines are AFLW test results.

with other results, because the author [2] applied the affine transformation on the face images before super-resolving. As shown in Fig. 6, by progressively self-supervised residual learning, our network generates the face images that are closer to the real face images than LapSRN [6] and DBPN [18]. The proposed network explicitly learns the edge map of high-resolution face images and embed it into the reconstruction of high-resolution face images to maintain the global structure of face images. So the face images generated by our RBPNet further reduce distortion than CBN [2] and URDGN [1]. As shown in Fig. 6(j), we accurately restored the global structure of the face images.

We also give efficiency test by calculating FLOPs and parameters. Considering the time cost of each method can be affected by many factors, such as code framework structure, the state of the GPU, running environment etc. So we estimate network performance and time consumption by calculating FLOPs and parameters. FLOPs is the number of floating point operations, which measure the complexity of an algorithm/model. Notice that FLOPs is

Please cite this article as: X. Chen, X. Wang and Y. Lu et al., RBPNET: An asymptotic Residual Back-Projection Network for super-resolution of very low-resolution face image, Neurocomputing, https://doi.org/10.1016/j.neucom.2019.09.079

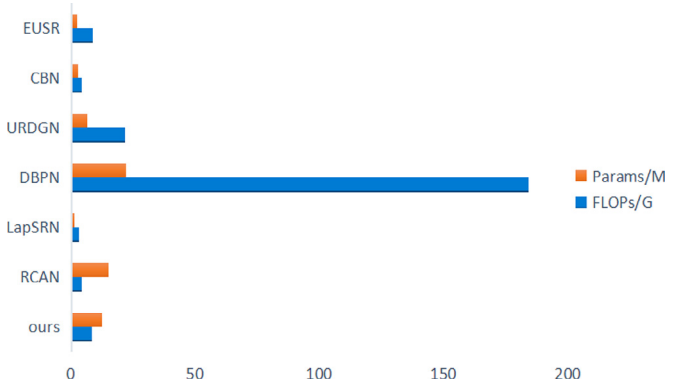

Fig. 7. The comparison of FLOPs and parameters among the state-of-the-arts.

different from FLOPS (the number of floating point operation per second). From Fig. 7, we can see that DBPN gets the largest value of evaluation index both parameters and FLOPs. Besides, the parameters and FLOPs of other methods are similar. LapSRN is trained on the Y-channel, so it has fewest parameters in all of compared methods. In addition to the above methods, our method get a comparable result compared with RCAN, EUSR, and URDGN in parameters and FLOPs.

5. Conclusion

We propose an asymptotic Residual Back-Projection Network (RBPNet) for super-resolution of very low-resolution face image. Unlike previous methods predicting high-resolution images by one-step learning, which tends to generate over-smoothed images, we propose to project the reconstructed high-resolution feature map back to the original low-resolution feature space to supervise reconstructed HR face images (self-supervision mechanism) and learn to map the residuals in low-resolution feature space to high-resolution feature space. We gradually approach the real high-resolution feature maps by multi-step residual learning. Extensive experiments demonstrate that our RBPNet achieves better performance than the state-of-the-art quantitatively and qualitatively.

Declaration of Competing Interest

The authors declared that they have no conflicts of interest to this work.

Acknowledgments

This work is supported by the National Natural Science Foundation of China (No. 61273273), by the National Key Research and Development plan (No. 2017YFC0112001), and by China Central Television (JG2018-0247).

References

[1] X. Yu, F. Porikli, Ultra-resolving face images by discriminative generative networks, in: Proceedings of the 14th European Conference on Computer Vision – ECCV 2016, 2016, pp. 318–333, doi:10.1007/978-3-319-46454-1_20. Amsterdam, The Netherlands
[2] S. Zhu, S. Liu, C.C. Loy, X. Tang, Deep cascaded bi-network for face hallucination, in: Proceedings of the 14th European Conference on Computer Vision – ECCV, 2016, pp. 614–630, doi:10.1007/978-3-319-46454-1_37. Amsterdam, The Netherlands
[3] C. Yang, S. Liu, M. Yang, Structured face hallucination, in: Proceedings of the 2013 IEEE Conference on Computer Vision and Pattern Recognition, 2013, pp. 1099–1106, doi:10.1109/CVPR.2013.146. Portland, OR, USA
[4] X. Yu, F. Porikli, Imagining the unimaginable faces by deconvolutional networks, IEEE Trans. Image Process. 27 (6) (2018) 2747–2761, doi:10.1109/TIP.2018.2808840.
[5] R. Dahl, M. Norouzi, J. Shlens, Pixel recursive super resolution, in: Proceedings of the IEEE International Conference on Computer Vision, ICCV 2017, 2017, pp. 5449–5458, doi:10.1109/ICCV.2017.581. Venice, Italy
[6] W. Lai, J. Huang, N. Ahuja, M. Yang, Deep Laplacian pyramid networks for fast and accurate super-resolution, in: Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, 2017, pp. 5835–5843. Honolulu, HI, USA
[7] M. Irani, S. Peleg, Improving resolution by image registration, CVGIP: Graph. Model Image Process. 53 (3) (1991) 231–239.
[8] S. Dai, M. Han, W. Xu, Y. Wu, Y. Gong, Soft edge smoothness prior for alpha channel super resolution, in: Proceedings of the 2007 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2007), 2007, doi:10.1109/CVPR.2007.383028. 18–23 June 2007
[9] Y. Tai, S. Liu, M.S. Brown, S. Lin, Super resolution using edge prior and single image detail synthesis, in: Proceedings of the Twenty-Third IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2010, 2010, pp. 2400–2407, doi:10.1109/CVPR.2010.5539933. San Francisco, CA, USA
[10] I.J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A.C. Courville, Y. Bengio, Generative adversarial networks, Advances in Neural Information Processing Systems 27 (2014) 2672–2680. arXiv:1406.2661.
[11] C. Dong, C.C. Loy, K. He, X. Tang, Image super-resolution using deep convolutional networks, in: Proceedings of the IEEE Transactions on Pattern Analysis and Machine Intelligence, 38, IEEE, 2016, pp. 295–307.
[12] W. Shi, J. Caballero, F. Caballero, J. Totz, A.P. Aitken, R. Bishop, D. Rueckert, Z. Wang, Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network, in: Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, 2016, pp. 1874–1883.
[13] S. Li, R. Fan, G. Lei, G. Yue, C. Hou, A two-channel convolutional neural network for image super-resolution, Neurocomputing 275 (2018) 267–277, doi:10.1016/j.neucom.2017.08.041.
[14] Z. Feng, J. Lai, X. Xie, J. Zhu, Image super-resolution via a densely connected recursive network, Neurocomputing 316 (2018) 270–276, doi:10.1016/j.neucom.2018.07.076.
[15] Y. Zhang, K. Li, K. Li, L. Wang, B. Zhong, Y. Fu, Image super-resolution using very deep residual channel attention networks, in: Proceedings of the 15th European Conference on Computer Vision – ECCV 2018, 2018, pp. 294–310, doi:10.1007/978-3-030-01234-2_18. Munich, Germany
[16] J. Kim, J. Lee, Deep residual network with enhanced upscaling module for super-resolution, in: Proceedings of the 2018 IEEE Conference on Computer Vision and Pattern Recognition Workshops, CVPR Workshops, 2018, pp. 800–808, doi:10.1109/CVPRW.2018.00124. http://openaccess.thecvf.com/content_cvpr_2018_workshops/w13/html/Kim_Deep_Residual_Network_CVPR_2018_paper.html. Salt Lake City, UT, USA
[17] Y. Zhao, R. Wang, W. Jia, W. Wang, W. Gao, Iterative projection reconstruction for fast and efficient image upsampling, Neurocomputing 226 (2017) 200–211, doi:10.1016/j.neucom.2016.11.049.
[18] M. Haris, G. Shakhnarovich, N. Ukita, Deep back-projection networks for super-resolution, in: Proceedings of the 2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, 2018, pp. 1664–1673. http://openaccess.thecvf.com/content_cvpr_2018/html/Haris_Deep_Back-Projection_Networks_CVPR_2018_paper.html. Salt Lake City, UT, USA
[19] C. Ledig, L. Theis, F. Huszar, J. Caballero, A. Cunningham, A. Acosta, A.P. Aitken, A. Tejani, J. Totz, Z. Wang, W. Shi, Photo-realistic single image super-resolution using a generative adversarial network, in: Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, 2017, pp. 105–114, doi:10.1109/CVPR.2017.19. Honolulu, HI, USA
[20] X. Xu, D. Sun, J. Pan, Y. Zhang, H. Pfister, M. Yang, Learning to super-resolve blurry face and text images, in: Proceedings of the IEEE International Conference on Computer Vision, ICCV 2017, 2017, pp. 251–260, doi:10.1109/ICCV.2017.36. Venice, Italy
[21] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, International Conference on Learning Representations (ICLR) (2015) arXiv:1409.1556.
[22] D.P. Kingma, J. Ba, Adam: a method for stochastic optimization, International Conference on Learning Representations (ICLR) (2015) arXiv:1412.6980.
[23] K. He, X. Zhang, S. Ren, J. Sun, Delving deep into rectifiers: Surpassing human-level performance on ImageNet classification, in: Proceedings of the 2015 IEEE International Conference on Computer Vision, ICCV 2015, 2015, pp. 1026–1034, doi:10.1109/ICCV.2015.123. Santiago, Chile
[24] Z. Liu, P. Luo, X. Wang, X. Tang, Deep learning face attributes in the wild, in: Proceedings of International Conference on Computer Vision (ICCV), 2015.
[25] V. Le, J. Brandt, Z. Lin, L.D. Bourdev, T.S. Huang, Interactive facial feature localization, in: Proceedings of the 12th European Conference on Computer Vision - ECCV 2012, 2012, pp. 679–692, doi:10.1007/978-3-642-33712-3_49. Florence, Italy
[26] P.M.R. Martin Koestinger, P. Wohlhart, H. Bischof, Annotated facial landmarks in the wild: a large-scale, real-world database for facial landmark localization, in: Proceedings of the First IEEE International Workshop on Benchmarking Facial Image Analysis Technologies, 2011.
[27] C. Ledig, L. Theis, F. HuszÃ!'r, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, Johannes, Z. Wang, W. Shi, Photo-realistic single image super-resolution using a generative adversarial network, in: Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 105–114.

Please cite this article as: X. Chen, X. Wang and Y. Lu et al., RBPNET: An asymptotic Residual Back-Projection Network for super-resolution of very low-resolution face image, Neurocomputing, https://doi.org/10.1016/j.neucom.2019.09.079

Xiaozhen Chen received the B.S. degree from Shenyang Institute of Engineering, in 2015. She is currently pursuing the M.E. degree in Biomedical Engineering at Beijing Institute of Technology, Beijing, China. Her main research interests include Image processing and pattern recognition.

Xuebo Wang received the B.S. degree from Taiyuan University of Technology, in 2017. He is currently pursuing the M.E. degree in Computer Science at Beijing Institute of Technology, Beijing, China. His main research interests include Image processing and pattern recognition.

Yao Lu received the B.S. degree in electronics from Northeast University, Shenyang, China, in 1982 and the Ph.D. degree in computer science from Gunma University, Gunma, Japan, in 2003. He was a Lecturer and an Associate Professor with Hebei University, China, from 1986 to 1998, and a foreign researcher with Gunma University in 1999. In 2003, he was an invited professor with the Engineering Faculty, Gunma Universitya Visiting Fellow of University of Sydney, Australia. He is currently a Professor with the Department of Computer Science, Beijing Institute of Technology, Beijing, China. He has published more than 100 papers in international conferences and journals. His research interests include neural network, image processing and video analysis, and pattern recognition.

Weiqi Li received the B.S. degree from Jilin University, in 2017. She is currently pursuing the M.E. degree in Computer Science at Beijing Institute of Technology, Beijing, China. Her main research interests include Image processing and pattern recognition.

Zijian Wang received the B.S. degree of animation technology from Communication University of China, in 2004, and M.E. degree of software engineering from Beijing University of Posts and Telecommunications, in 2011. He is engaged in computer visual effect development in China Central Television. His research interests include image processing and pattern recognition.

Zhuowei Huang received the B.S. degree in Mechanical and Electronics from Beijing Institute of Technology, Beijing, China, in 2005 and the Master's degree in software engineering from Beijing Institute of Technology, China, in 2008. He got another Master's degree in management from Tsinghua University, Beijing, China, in 2016. He is now a senior engineer with CCTV (China Central Television) and the main builder of the CCTV Big Data Platform. His research interests include Big data and Artificial intelligence.

Please cite this article as: X. Chen, X. Wang and Y. Lu et al., RBPNET: An asymptotic Residual Back-Projection Network for super-resolution of very low-resolution face image, Neurocomputing, https://doi.org/10.1016/j.neucom.2019.09.079