Wavelet-SRNet: A Wavelet-based CNN for Multi-scale Face Super Resolution

Huaibo Huang[1,2,3], Ran He[1,2,3], Zhenan Sun[1,2,3] and Tieniu Tan[1,2,3]
[1]School of Engineering Science, University of Chinese Academy of Sciences
[2]Center for Research on Intelligent Perception and Computing, CASIA
[3]National Laboratory of Pattern Recognition, CASIA
huaibo.huang@cripac.ia.ac.cn, {rhe, znsun, tnt}@nlpr.ia.ac.cn

Abstract

Most modern face super-resolution methods resort to convolutional neural networks (CNN) to infer high-resolution (HR) face images. When dealing with very low resolution (LR) images, the performance of these CNN based methods greatly degrades. Meanwhile, these methods tend to produce over-smoothed outputs and miss some textural details. To address these challenges, this paper presents a wavelet-based CNN approach that can ultra-resolve a very low resolution face image of 16×16 *or smaller pixel-size to its larger version of multiple scaling factors (*$2\times$*,* $4\times$*,* $8\times$ *and even* $16\times$*) in a unified framework. Different from conventional CNN methods directly inferring HR images, our approach firstly learns to predict the LR's corresponding series of HR's wavelet coefficients before reconstructing HR images from them. To capture both global topology information and local texture details of human faces, we present a flexible and extensible convolutional neural network with three types of loss: wavelet prediction loss, texture loss and full-image loss. Extensive experiments demonstrate that the proposed approach achieves more appealing results both quantitatively and qualitatively than state-of-the-art super-resolution methods.*

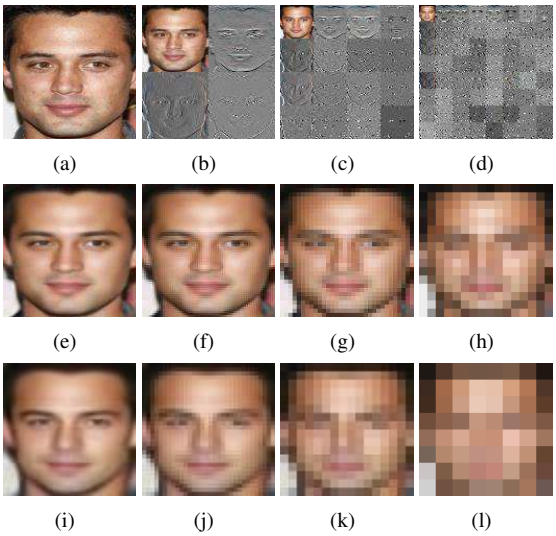


Figure 1. Illustration of wavelet decomposition and our wavelet-based SR. Top row: (a) The original 128×128 high-resolution face image and its (b) 1 level, (c) 2 level, (d) 3 level, full wavelet packet decomposition image. Middle row: (h) The 16×16 low-resolution face image and its (g) $2\times$, (f) $4\times$, (e) $8\times$, upscaling versions inferred by our network. Bottom row: similar with the middle row except the low-resolution input (l) is 8×8 pixel-size.

1. Introduction

Face super-resolution (SR), also known as face hallucination, refers to reconstructing high resolution (HR) face images from their corresponding low resolution (LR) inputs. It is significant for most face-related applications, e.g. face recognition, where captured faces are of low resolution and lack in essential facial details. It is a special case of single image super resolution and many methods have been proposed to address it. It is a widely known undetermined inverse problem, i.e., there are various corresponding high-resolution answers to explain a given low-resolution input.

Most current single image super-resolution methods [2, 6, 14, 15, 23] depend on a pixel-wise mean squared error (MSE) loss in image space to push the outputs pixel-wise closer to the ground-truth HR images in training phase. However, such approaches tend to produce blurry and over-smoothed outputs, lacking some textural details. Besides, they seem to only work well on limited up-scaling factors ($2\times$ or $4\times$) and degrades greatly when ultra-resolving a very small input (like 16×16 or smaller). Several recent efforts [5, 33, 35] have been developed to deal with this issue based on convolutional neural networks. Dahl et al. [5] use PixelCNN [27] to synthesize realistic details. Yu et al. [33] investigate GAN [8] to create perceptually realistic results. Zhu et al. [35] combine dense correspondence field estimation with face super-resolution. However, the application of these methods in super-resolution in image space faces

many problems, such as computational complexity [5], instability in training [33] and poor robustness for pose and occlusion variations [35]. Therefore, due to various problems yet to be solved, image SR remains an open and challenging task.

Wavelet transform (WT) has been shown to be an efficient and highly intuitive tool to represent and store multi-resolution images [18]. It can depict the contextual and textural information of an image at different levels, which motivates us to introduce WT to a CNN-based super-resolution system. As illustrated in Figure 1, the approximation coefficients(i.e. the top-left patches in (b-d)) of different-level wavelet packet decomposition [4] compress the face's global topology information at different levels; the detail coefficients(i.e. the rest patches in (b-d)) reveal the face's structure and texture information. We assume that a high-quality HR image with abundant textural details and global topology information can be reconstructed via a LR image as long as the corresponding wavelet coefficients are accurately predicted. Hence, the task of inferring a high-resolution face is transformed to predicting a series of wavelet coefficients. Emphasis on the prediction of high-frequency wavelet coefficients helps recovering texture details, while constraints on the reconstruction of low-frequency wavelet coefficients enforces consistence on global topology information. The combination of the two aspects makes the final HR results more photo-realistic.

To take full advantage of wavelet transform, we present a wavelet-based convolutional neural network for face super-resolution, which consists of three subnetworks: embedding, wavelet prediction and reconstruction networks. The embedding net takes the low-resolution face as an input and represents it as a set of feature maps. The wavelet prediction net is a series of parallel individual subnetworks, each of which aims to learn a certain wavelet coefficient using the embedded features. The number of these subnetworks is flexible and easy to adjust on demand, which makes the magnification factor flexible as well. The reconstruction network is used to recover the inferred wavelet coefficients to the expected HR image, acting as a learned matrix. These three subnetworks are coordinated with three types of loss: wavelet prediction loss, texture loss and full-image loss. The wavelet prediction loss and texture loss correspond with the wavelet prediction subnetwork, imposing constraint in wavelet domain. The full-image loss is used after the reconstruction subnetwork to add a traditional MSE constraint in image space. Besides, as each wavelet coefficient shares the same size with the low-resolution input, we use a network configuration to make every feature map keep the same size with the input, which reduces the difficulty of training. As our network is fully convolutional and trained with simply-aligned faces, it can apply to different input resolutions with various magnifications, regardless of pose and occlusion variations. Experimental results collaborate with our assumption and demonstrate that our method can well capture both global topology information and local textural details of human faces.

Main contributions of our work can be summarized as follows:

1) A novel wavelet-based approach is proposed for CNN-based face SR problem. To the best of our knowledge, this is the first attempt to transform single image SR to wavelet coefficients prediction task in deep learning framework - albeit many wavelet-based researches exist for SR.

2) A flexible and extensible fully convolutional neural network is presented to make the best use of wavelet transform. It can apply to different input-resolution faces with multiple upscaling factors.

3) We qualitatively and quantitatively explore multi-scale face super-resolution, especially on very low input resolutions. Experimental results show that our proposed approach outperforms state-of-the-art face SR methods.

2. Related work

In general, image super-resolution methods can be divided into three types: interpolation-based, statistics-based [26, 31, 32] and learning-based methods [3, 9, 24]. In the early years, the former two types have attracted most of attention for their computationally efficiency. However, they are always limited to small upscaling factors. Learning based methods employ large quantities of LR/HR image pair data to infer missing high-frequency information and promises to break the limitations of big magnification. Recently deep learning based methods [6, 14, 15, 2, 23] have been introduced into SR problem due to their powerful ability to learn knowledge from large database. Most of these convolutional methods use MSE loss to learn the map function of LR/HR image pairs, which leads to over-smooth outputs when the input resolution is very low and the magnification is large.

Specific to face super-resolution, there have been about three ways to alleviate this problem. The first one [29, 13, 28, 30, 35] is to exploits the specific static information of face images with the help of face analysis technique. Yang et al. [29] estimate landmarks and facial pose before reconstructing HR images while the accurate estimation is difficult for rather small faces. Zhu et al. [35] present a unified framework of face super-resolution and dense correspondence field estimation to recover textural details. They achieve state-of-the-art results for very low resolution inputs but fail on faces with various poses and occlusions, due to the difficulty of accurate spatial prediction.

The second way [17, 33, 25, 5] is to bring in image prior knowledge with the help of generative models. Yu et al. [33] propose a generative adversarial network (GAN [8]) to re-

solve 16×16 pixel-size faces to its $8\times$ larger versions. Dahl et al. [5] present a recursive framework based on PixelCNN [27] to synthesize details of $4\times$ magnified images with 8×8 LR inputs. The 32×32 outputs are not sufficiently perceptual appealing, and their method suffers from high computational complexity.

The last way is to introduce perceptual losses to improve the outputs' perceptual quality directly. Johnson et al. [12] use feature reconstruction loss as perceptual loss to recover more semantic information. However, reconstruction features are not as intuitive as wavelet coefficients to depict perceptual quality.

Many wavelet-based methods have already been proposed for super resolution problem. A large percentage of them focus on multiple images SR [22, 10], which means inferring a high-resolution image from a sequence of low-resolution images. As for single image super resolution, wavelet transform is mostly used to help interpolation-based [1, 21] and static-based [34] methods. Naik et al. [21] propose a modified version of classical wavelet-based interpolation method [1]. Gao et al. [7] propose a hybrid wavelet convolution network. They use wavelet to provide a set of sparse coding candidates and another convolution net for sparse coding, which is totally different with ours. Besides, Mallat [19] uses wavelet transform to separate the variations of data at different scales, while we predict the wavelets from LR inputs, designed especially for super resolution.

3. Approach

In this section, we present a novel framework for face super resolution, which predicts a series of corresponding wavelet coefficients instead of HR images directly. Special losses in wavelet domain are designed to capture both global topology information and local textural details. Then, an extensible fully convolutional neural network (Wavelet-SRNet) is proposed for multi-scale face super resolution. At last, implement details of Wavelet-SRNet are given.

3.1. Wavelet packet transform

Our method is based on wavelet transform, more specifically wavelet packet transform (WPT), which decomposes an image into a sequence of wavelet coefficients of the same size. We choose the simplest wavelet, Haar wavelet, for it is enough to depict different-frequency facial information. We use 2-D fast wavelet transform (FWT) [20] to compute Haar wavelets. The wavelet coefficients at different levels are computed by repeating the decomposition in Figure 2 to each output coefficient iteratively. Example results of WPT is showed in Figure 1 (b-d).

3.2. Wavelet-based super resolution

Generic single image super resolution aims to learn a map function $f_\theta(x)$ defined by the parameter θ to estimate a high resolution image $\hat{y}$ with a given low resolution input x. Suppose that y denotes a ground-truth HR image and $D \equiv \{(x_i, y_i)\}_i^N$ represents a large dataset of LR/HR image pairs, then most current learning-based SR methods optimize the parameter θ through the following form

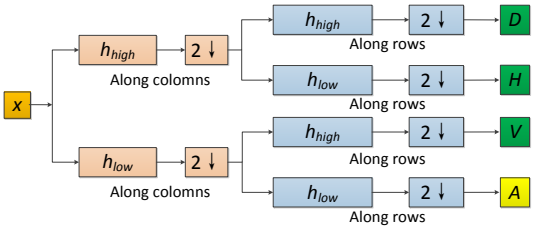


Figure 2. Illustration of fast wavelet transform (FWT). FWT uses low-pass and high-pass decomposition fillers iteratively to compute wavelet coefficients, where Haar-based $h_{low} = (1/\sqrt{2}, 1/\sqrt{2})$ and $h_{high} = (1/\sqrt{2}, -1/\sqrt{2})$.

$$\arg\max_{\theta} \sum_{(x,y)\in D} \log p(y|x) \tag{1}$$

The most common loss function is pixel-wise MSE in HR image space

$$l_{mse}(\hat{y}, y) = \|\hat{y} - y\|_F^2 \tag{2}$$

As argued in many papers [17, 33, 25, 5], merely minimizing MSE loss can hardly capture high-frequency texture details to produce satisfactory perceptual results. As texture details can be depicted by high-frequency wavelet coefficients, we transform super resolution problem from original image space to wavelet domain and introduce wavelet-based losses to help texture reconstruction.

Consider n-level full wavelet packet decomposition, where n determines the scaling factor r of super resolution and the number of wavelet coefficients N_w, i.e., $r = 2^n$, $N_w = 4^n$. Let $C = (c_1, c_2, \cdots, c_{N_w})$ and $\hat{C} = (\hat{c}_1, \hat{c}_2, \cdots, \hat{c}_{N_w})$ denote the ground-truth and inferred wavelet coefficients, the model parameter θ of the map function $g_\theta(x) = (g_{\theta,1}(x), g_{\theta,1}(x), \cdots, g_{\theta,N_w}(x))$ can be optimized by the form

$$\arg\max_{\theta} \sum_{(x,C)\in D} \log p(C|x) \tag{3}$$

We propose two kinds of wavelet-based loss: wavelet prediction loss and texture loss. The former one is a weighted version of MSE in wavelet domain, defined as

$$\begin{aligned} l_{wavelet}(\hat{C}, C) &= \|W^{1/2} \odot (\hat{C} - C)\|_F^2 \\ &= \sum_{i=1}^{N_w} \lambda_i \|\hat{c}_i - c_i\|_F^2 \\ &= \lambda_1 \|\hat{c}_1 - c_1\|_F^2 + \sum_{i=2}^{N_w} \lambda_i \|\hat{c}_i - c_i\|_F^2 \end{aligned} \tag{4}$$

where $W = (\lambda_1, \lambda_2, \cdots, \lambda_{N_w})$ is the weight matrix to balance the importance of different-band wavelet coefficients. More attention can be paid on local textures with bigger weights appointed to high-frequency coefficients. Meanwhile, the term $\|\hat{c}_1 - c_1\|_F^2$ captures global topology information and serves as the loss function of an auto-encoder when the approximation coefficient c_1 is taken as input, which is helpful for maintaining training stability.

The texture loss is designed to prevent high-frequency wavelet coefficients from converging to zero, defined as

$$l_{texture} = \sum_{i=k}^{N_w} \gamma_i \max(\alpha \|c_i\|_F^2 + \varepsilon - \|\hat{c}_i\|_F^2, 0) \quad (5)$$

where k indicates the start index of the wavelet coefficients to be penalized for taking small values, γ_i is balance weights, α and ε are slack values. It keeps high-frequency wavelet coefficients non-zero and hence prevents the degradation of texture details.

A traditional MSE loss in image space, which is called full-image loss, is also used to get a balance between smoothness and textures. The unified loss function is defined as follows

$$\begin{aligned} l_{total} &= l_{wavelet} + \mu l_{texture} + \nu l_{full-image} \\ &= \sum_{i=1}^{N_w} \lambda_i \|\hat{c}_i - c_i\|_F^2 \\ &+ \mu \sum_{i=k}^{N_w} \gamma_i \max(\alpha \|c_i\|_F^2 + \varepsilon - \|\hat{c}_i\|_F^2, 0) \\ &+ \nu \|R\hat{C} - y\|_F^2 \end{aligned} \quad (6)$$

where μ and ν are the balance parameters, and R is the reconstruction matrix to generate $\hat{y}$ from $\hat{C}$, i.e., $\hat{y} = R\hat{C}$.

3.3. Network Architecture

As outlined in Figure 3, our wavelet-based convolutional neural network consists of three subnetworks: embedding, wavelet prediction, reconstruction networks. The embedding net represents the low-resolution input as a set of feature maps. Then the wavelet prediction net estimates the corresponding wavelet coefficient images. Finally the reconstruction net reconstructs the high-resolution image from these coefficient images.

The **embedding net** takes a low-resolution image of the size $3 \times h \times w$ as input and represents it as a set of feature maps. All the convolution fillers share the same size of 3×3 with a stride of 1 and a pad of 1, which makes every feature map in the embedding net the same size with the input image. The number of feature maps (or the channel-size) increases in the forward direction to explore enough information for wavelet prediction. Through the embedding net, the input LR image is mapped to feature maps of the size $N_e \times h \times w$ without up-sampling or down-sampling, where N_e is the last layer's channel-size.

The **wavelet prediction net** can be split into N_w parallel independent subnets, where $N_w = 4^n$ on the condition that the level of wavelet-packet decomposition is n and the magnification $r = 2^n$. Each of these subnets takes the output feature maps of the embedding net as input and generates the corresponding wavelet coefficient. We set all the convolution fillers the size of 3×3 with a stride of 1 and a pad of 1 similarly with the embedding net, so that every inferred wavelet coefficient is the same size with the LR input, i.e, $3 \times h \times w$. Besides, motivated by the high independence between the coefficients of Haar wavelet transform, no information is allowed to interflow between every two subnets, which makes our network extensible. It is easy to realize different magnifications with different numbers of the subnets in the prediction net. For example, $N_w = 16$ and $N_w = 64$ stand for $4\times$ and $8\times$ magnifications, respectively.

The **reconstruction net** is used to transform the wavelet images of the total size $N_w \times 3 \times h \times w$ into the original image space of the size $3 \times (r \times h) \times (r \times w)$. It comprises a deconvolution layer with a filler size of $r \times r$ and a stride of r. Although the size of the deconvolution layer is dependent on the magnification r, it can be initialized by a constant wavelet reconstruction matrix and fixed in training. Hence it has no effect on the extensibility of the whole networks.

As mentioned above, all the convolution fillers of the embedding and wavelet prediction nets share the same size of 3×3 with a stride of 1 and a pad of 1, keeping every feature map the same spatial size with the input image. This reduces both the size of model parameters and the computation complexity. Besides, to prevent gradients exploding/vanishing and accelerate convergence, we use skip-connections between every two layers except the first layer. Batch-norm is also used after every layer.

The definition of our networks can be formulated as follows

$$\begin{aligned} \hat{y} &= \phi(\hat{C}) = \phi\{(\hat{c}_1, \hat{c}_2, \cdots, \hat{c}_{N_w})\} \\ &= \phi\{(\varphi_1(\hat{z}), \varphi_2(\hat{z}), \cdots, \varphi_{N_w}(\hat{z}))\} \\ &= \phi\{(\varphi_1(\psi(x)), \varphi_2(\psi(x)), \cdots, \varphi_{N_w}(\psi(x)))\} \end{aligned} \quad (7)$$

where

$$\begin{aligned} &\psi : R^{3\times h\times w} \to R^{N_e \times h \times w} \\ &\varphi_i : R^{N_e \times h \times w} \to R^{3\times h\times w}, i = 1, 2, \cdots, N_w \\ &\phi : R^{N_w \times 3 \times h \times w} \to R^{3\times (r\times h)\times (r\times w)} \end{aligned} \quad (8)$$

are mappings of the embedding,wavelet prediction, reconstruction nets, respectively.

3.4. Implementation details

A novel training technique for face super-resolution, called as **co-training**, is used to make our model stable in

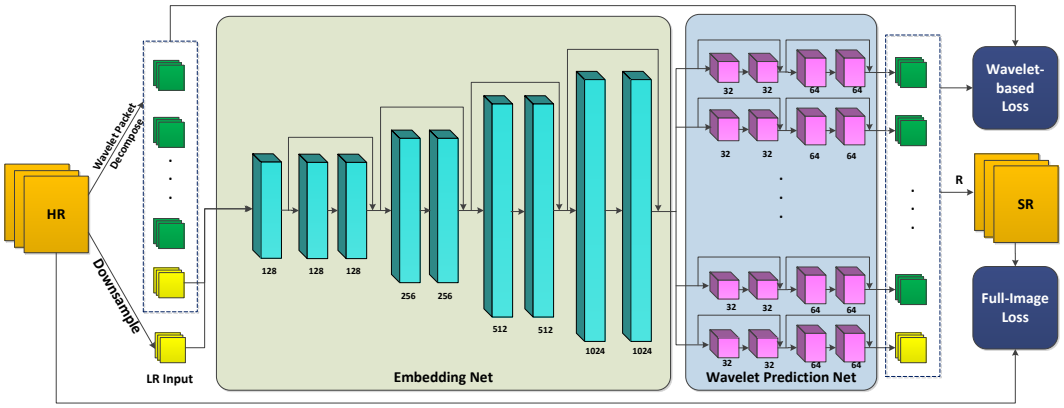


Figure 3. The architecture of our wavelet-based super-resolution net (Wavelet-SRNet). All the convolution layers have the same filter map size of 3x3 and each number below them defines their individual channel size. Skip connections exist between every two convolution layers (except the first layer) in the embedding and wavelet predicting nets.

training and robust toward unknown down-sampling methods. Two types of low-resolution images are taken as input, one of which is down-sampled by bicubic interpolation and the other is the approximation coefficient of wavelet packet decomposition. Take the case of 16×16 input-resolution resolved to 128×128 for example. We resize all center-cropped faces to 134×134 with bicubic interpolation and randomly crop them to 128×128 images. Wavelet packet decomposition at 3 level is used to get the ground-truth wavelet coefficients c_i in (6). The approximation coefficient c_1 is treated as one version of low-resolution input. With the mapping function $\hat{c_1} = \varphi_1(\psi(c_1))$, and the distance constraint $\|\hat{c_1} - c_1\|_F^2$, the embedding and prediction nets serve as an auto-encoder, which assures no loss of the original input information and facilitates training stability. Another version of low-resolution input, directly down-sampled by bicubic interpolation, is used cooperatively with the wavelet version, which helps maintaining the robustness of our model. In the testing phase, we evaluate on faces down-sampled by bicubic interpolation.

Since our network is a fully convolutional architecture without fully-connected layers, it can be applied to the input of arbitrary size. We firstly train a model for 16×16 input resolution with $8\times$ magnification, and then fine-tune it for 8×8 input resolution with $8\times$ magnification. For 8×8 input resolution with $16\times$ magnification, we initialize the parameters by the overlapping ones of the model for 8×8 with $8\times$ magnification before fine-tuning it. For other cases, we just choose the closest model for evaluation.

Our model is implemented with the Caffe framework [11]. The loss in (6) is minimized using SGD with a batch size of 256. For the hyper-parameters, we set empirically $\lambda_1 = 0.01, \lambda_2 = \lambda_3 = \cdots = \lambda_{N_w} = 1, \mu = 1, k = 2, \gamma_k = \gamma_{k+1} = \cdots = \gamma_{N_w} = 1, \nu = 0.1$. The learning rate is set to 0.01 initially and reduced by a factor of 10 each 3000 iterations. It takes about $14,000 \sim 16,000$ iterations for our network to converge.

4. Experiments

Our experiments are implemented on two datasets: CelebA [36] and Helen [16]. There are 202,599 faces in CelebA and 2,230 faces in Helen. In the training phase, we use the large train set of CelebA(162,700 images) for training and the validation set(19,867 images) of CelebA for validation. In the testing phase, we evaluate with the 19,962-image test set of CelebA and the 330-image test set of Helen, assuring no over-lapped images appearing in both the training and testing phase. The images are cropped around the face with eyes aligned horizontally.

We evaluate the performance of Wavelet-SRNet on multiple input resolutions, comparing with bicubic interpolation, wavelet-based interpolation (WaveletIP, for short) [21] and state-of-the-art methods: SRCNN [6], URDGN [33], CBN [35]. WaveletIP [21] upsample images in both spatial and wavelet domain. SRCNN [6] is a generic cnn-based super resolution method so we retrain it on CelebA training set to suit better for face images. URDGN [33] and CBN [35] are trained on CelebA. URDGN chooses 15,000 and 500 images randomly from CelebA for training and evaluation respectively. CBN uses the whole CelebA dataset for training. Hence their results on Helen may be more persuasive than on the CelebA test set. For a fair comparison, we use the same set of eyes-aligned faces for all the methods with no extra preprocessing before down-sampling. We adopt PSNR(dB) and SSIM for quantitative metric, and calculate PSNR on the luminance channel, following by [35], and SSIM on the three channels of RGB.

4.1. Results on multiple resolutions

As mentioned above, our method can apply to different input resolutions with multiple magnifications. In Figure 4, we show the qualitative results of our method on different input resolutions comparing with the bicubic interpolation baseline. Our method can reconstruct faces from very small inputs of 8×8 pixel-size and the inferred outputs are perceptually identity-persistent to some degree, which implies that a small number of 64 pixels constains most of a face's identity information. Besides, while the outputs of 8×8 input resolution are still a little blurry, the outputs of the larger input resolutions are very close to the original high resolution faces in human perception.

4.2. Comparison on very low resolutions

We compare our method qualitatively with state-of-the-art methods on two very low resolution cases, 16×16 and 8×8, both with a magnification of 8.

As for 16×16 input resolution in Figure 5 (a-h), our method achieves the best perceptual performance. Bicubic interpolation, WaveletIP [21] and SRCNN [6] fail to infer high-frequency details and generate over-smoothed outputs. URDGN [33] promises to predict high-frequency information with an adversarial loss. However, we evaluate URDGN with their offered model and find texture details over-synthesized, as Figure 5 (f) illustrates. It is perhaps because that their train and test sets are much smaller than ours and their adversarial networks lack in robustness. CBN [35] achieves the second-place performance except deformation in some cases. They hallucinate faces with the help of dense correspondence field estimation and consequently encounter abnormal results when unable to estimate facial locations accurately. Comparing with other methods, our network infers the high-frequency details directly in wavelet domain and the results prove its effectiveness.

As for 8×8 input resolution in Figure 5 (i-p), only our method, URDGN [33] and CBN [35] can reconstruct faces. While URDGN [33] contains much weird textures and CBN [35] tends to generate faces closer to the mean face, our results are more identity-similar to the ground-truth and plausible for human vision.

4.3. Discussion of the robustness

We evaluate the robustness of our method toward unknown Gaussian blur, poses and occlusions. In this section, we still adopt the same model used above, with no extra efforts to deal with these special cases.

In Figure 6, the low-resolution faces are generated by a Gaussian blur kernel with a stride of 8 corresponding to $8\times$ down-sampling. σ for Gaussian blur kernel increases from 0 to 6, where $\sigma = 0$ means nearest-neighbor interpolation down-sampling. As shown in Figure 6, our method demonstrates certain robustness when $\sigma < 4$ and generates smoother faces when $\sigma >= 4$. As a comparison, the results of CBN become more similar with mean face.

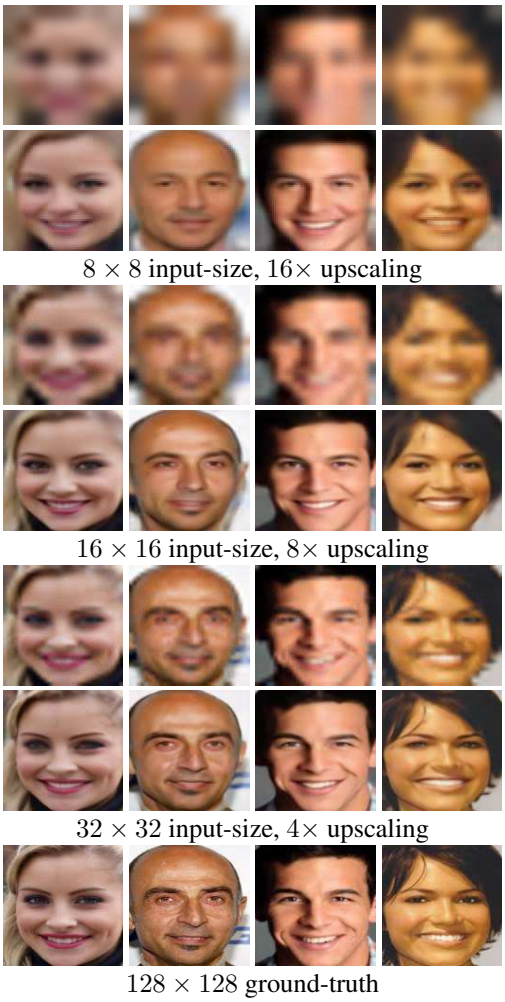


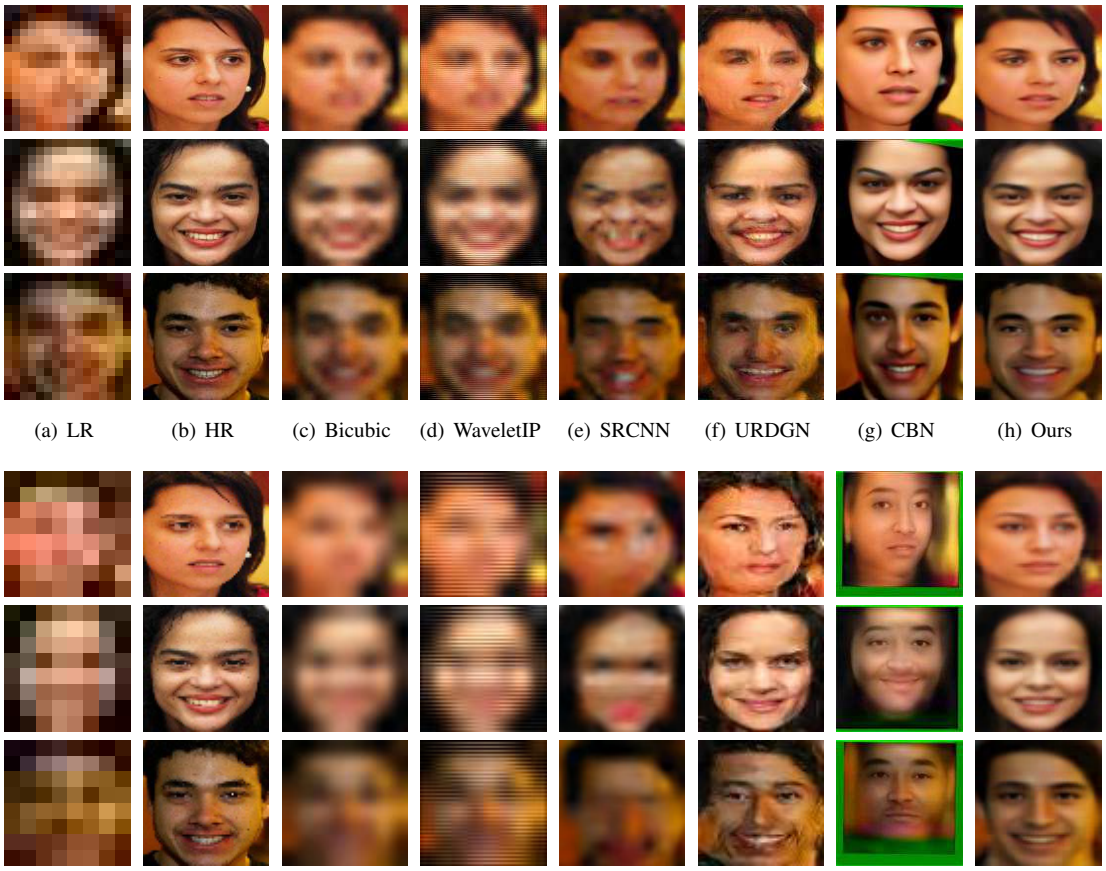
Figure 4. Results of various input resolutions: 8×8, 16×16 and 32×32. For each input resolution, the first row is generated by bicubic interpolation and the second is ours.

For pose variations, as shown in Figure 7, CBN fails to reconstruct plausible faces of large poses, perhaps due to inaccurate spatial prediction. Meanwhile, our method can still infer high-quality results.

For occlusion variations, we take some faces with natural occlusions for example. As shown in Figure 8, CBN tends to over-synthesize occluded facial parts, e.g., the eyes and lips, while ours resolves the occluded parts and the rest dependently.

4.4. Quantitative results

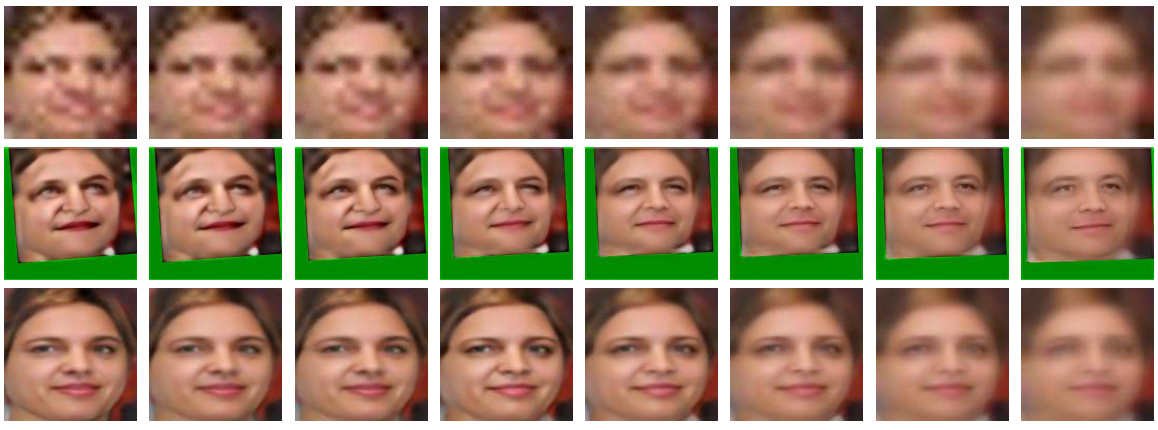
We evaluate Wavelet-SRNet quantitatively using average PSNR(dB) and SSIM on the two test sets of CelebA and Helen. We conduct evaluation on four cases: $(32 \times 32, 4\times)$,$(16 \times 16, 8\times)$,$(8 \times 8, 8\times)$ and $(8 \times 8, 16\times)$ ($(m \times m, n\times)$ means $m \times m$ input resolution with magni-



(a) LR (b) HR (c) Bicubic (d) WaveletIP (e) SRCNN (f) URDGN (g) CBN (h) Ours

(i) LR (j) HR (k) Bicubic (l) WaveletIP (m) SRCNN (n) URDGN (o) CBN (p) Ours

Figure 5. Comparison with state-of-the-art methods on very low input resolutions. The input resolutions are 16×16 and 8×8 for the top three and bottom three rows, respectively. The magnifications are both $8\times$. Images are selected randomly from Helen test set. We do not try to crop the green area caused by the shape transform of CBN in (g) and (o) to avoid facial deformation.



(a) $\sigma = 0$ (b) $\sigma = 0.5$ (c) $\sigma = 1$ (d) $\sigma = 2$ (e) $\sigma = 3$ (f) $\sigma = 4$ (g) $\sigma = 5$ (h) $\sigma = 6$

Figure 6. Robustness toward unknown gaussian blur on Helen test set. The input resolution is 16×16 and the magnification is $8\times$. The top, middle and bottom rows are the results of bicubic interpolation, CBN and ours, respectively.

fication of n). As is shown in Table 1, our method achieves the best quantitative performance. As expected, bicubic interpolation is better than other state-of-the-art methods because it is designed to minimize the pixel-wise MSE

Dataset	Method	32×32, $4\times$		16×16, $8\times$		8×8, $8\times$		8×8, $16\times$	
		PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
Celeba	Bicubic	29.20	0.9258	24.83	0.8525	21.84	0.7687	21.36	0.7838
	WPSR	25.01	0.8467	21.50	0.7234	19.44	0.6476	19.03	0.6332
	SRCNN	20.61	0.8004	20.15	0.7954	17.84	0.6927	18.39	0.6880
	URDGN	-	-	24.63	0.8527	21.41	0.7614	-	-
	CBN	25.93	0.8749	24.68	0.8369	19.93	0.7201	19.78	0.7327
	Ours	**30.56**	**0.9432**	**26.61**	**0.8949**	**23.35**	**0.8370**	**22.65**	**0.8201**
Helen	Bicubic	27.44	0.8762	23.96	0.7916	21.12	0.7068	20.96	0.7084
	WPSR	24.17	0.7845	21.10	0.6494	19.45	0.5881	18.85	0.5580
	SRCNN	21.93	0.8227	19.814	0.7321	17.46	0.6353	18.51	0.7367
	URDGN	-	-	23.12	0.7708	19.32	0.6416	-	-
	CBN	23.39	0.7773	22.44	0.7486	19.58	0.6301	19.78	0.7201
	Ours	**27.94**	**0.8827**	**24.63**	**0.8276**	**21.83**	**0.7662**	**21.80**	**0.7491**

Table 1. Quantitative results on CelebA and Helen test sets.

loss without considering the characteristics of human face. The results in Table 1 demonstrate the fact that our method preserves the pixel-wise consistence between LR inputs and HR ground-truth while generates perceptually plausible faces.

5. Conclusion

We propose a novel wavelet-based approach for multi-scale face super resolution, which transforms single image super resolution to wavelet coefficients prediction task in deep learning framework. A flexible wavelet-based convolutional neural network (Wavelet-SRNet) is presented, which consists of three subnetworks: embedding, wavelet prediction and reconstruction networks. Three types of loss, wavelet prediction loss, texture loss and full-image loss, are designed to capture both the global topology information and local texture information of human faces. Due to its extensible fully convolutional architecture trained with simply-aligned faces, our network is applicable to different input resolutions with various magnifications. Experimental results show that our method demonstrates promising robustness toward unknown Gaussian blur, poses and occlusions, and achieves better performance both qualitatively and quantitatively than the state-of-the-art.

Acknowledgement

This work is partially funded by National Natural Science Foundation of China (Grant No. 61622310, 61473289), the State Key Development Program (Grant No. 2016YFB1001001) and Beijing Municipal Science and Technology Commission (No.Z161100000216144).

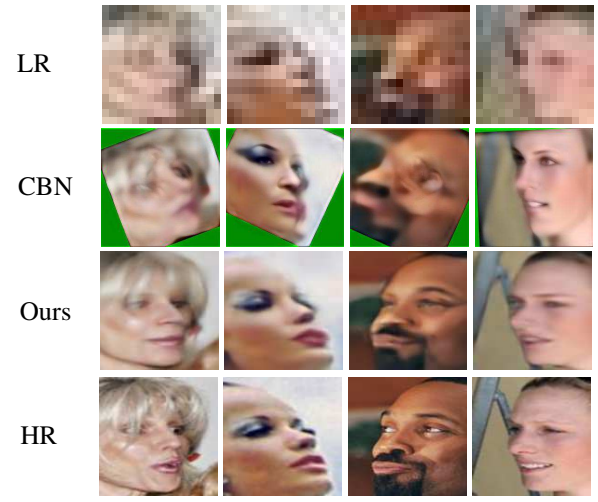


Figure 7. Results of 16×16 faces with large pose variations.

Figure 8. Results of 16×16 faces with occlusion variations.

References

[1] G. Anbarjafari and H. Demirel. Image super resolution based on interpolation of wavelet domain high frequency subbands and the spatial domain input image. *Etri Journal*, 32(3):390–394, 2010.

[2] J. Bruna, P. Sprechmann, and Y. LeCun. Super-resolution with deep convolutional sufficient statistics. *ICLR*, 2016.

[3] H. Chang, D.-Y. Yeung, and Y. Xiong. Super-resolution through neighbor embedding. In *CVPR*, volume 1, pages 275–282, 2004.

[4] R. R. Coifman and M. V. Wickerhauser. Entropy-based algorithms for best basis selection. *IEEE Transactions on Information Theory*, 38(2):713–718, 1992.

[5] R. Dahl, M. Norouzi, and J. Shlens. Pixel recursive super resolution. *arXiv preprint arXiv:1702.00783*, 2017.

[6] C. Dong, C. C. Loy, K. He, and X. Tang. Image super-resolution using deep convolutional networks. *TPAMI*, 38(2):295–307, 2016.

[7] X. Gao and H. Xiong. A hybrid wavelet convolution network with sparse-coding for image super-resolution. In *ICIP*, pages 1439–1443, 2016.

[8] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In *NIPS*, pages 2672–2680, 2014.

[9] J.-B. Huang, A. Singh, and N. Ahuja. Single image super-resolution from transformed self-exemplars. In *CVPR*, pages 5197–5206, 2015.

[10] H. Ji and C. Fermüller. Robust wavelet-based super-resolution reconstruction: theory and algorithm. *TPAMI*, 31(4):649–660, 2009.

[11] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell. Caffe: Convolutional architecture for fast feature embedding. *arXiv preprint arXiv:1408.5093*, 2014.

[12] J. Johnson, A. Alahi, and L. Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *ECCV*, pages 694–711, 2016.

[13] C. Jung, L. Jiao, B. Liu, and M. Gong. Position-patch based face hallucination using convex optimization. *IEEE Signal Processing Letters*, 18(6):367–370, 2011.

[14] J. Kim, J. Kwon Lee, and K. Mu Lee. Accurate image super-resolution using very deep convolutional networks. In *CVPR*, 2016.

[15] J. Kim, J. Kwon Lee, and K. Mu Lee. Deeply-recursive convolutional network for image super-resolution. In *CVPR*, pages 1637–1645, 2016.

[16] V. Le, J. Brandt, Z. Lin, L. Bourdev, and T. S. Huang. Interactive facial feature localization. In *ECCV*, pages 679–692, 2012.

[17] C. Ledig, L. Theis, F. Huszár, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, Z. Wang, et al. Photo-realistic single image super-resolution using a generative adversarial network. *arXiv preprint arXiv:1609.04802*, 2016.

[18] S. Mallat. Wavelets for a vision. *Proceedings of the IEEE*, 84(4):604–614, 1996.

[19] S. Mallat. Understanding deep convolutional networks. *Phil. Trans. R. Soc. A*, 374(2065):20150203, 2016.

[20] S. G. Mallat. A theory for multiresolution signal decomposition: the wavelet representation. *TPAMI*, 11(7):674–693, 1989.

[21] S. Naik and N. Patel. Single image super resolution in spatial and wavelet domain. *The International Journal of Multimedia & Its Applications*, 5(4):23, 2013.

[22] N. Nguyen and P. Milanfar. A wavelet-based interpolation-restoration method for superresolution (wavelet superresolution). *Circuits, Systems, and Signal Processing*, 19(4):321–338, 2000.

[23] W. Shi, J. Caballero, F. Huszar, J. Totz, A. P. Aitken, R. Bishop, D. Rueckert, and Z. Wang. Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network. In *CVPR*, 2016.

[24] A. Singh, F. Porikli, and N. Ahuja. Super-resolving noisy images. In *CVPR*, 2014.

[25] C. K. Sønderby, J. Caballero, L. Theis, W. Shi, and F. Huszár. Amortised map inference for image super-resolution. *ICLR*, 2017.

[26] J. Sun, Z. Xu, and H.-Y. Shum. Image super-resolution using gradient profile prior. In *CVPR*, pages 1–8, 2008.

[27] A. van den Oord, N. Kalchbrenner, L. Espeholt, O. Vinyals, A. Graves, et al. Conditional image generation with pixelCNN decoders. In *NIPS*, pages 4790–4798, 2016.

[28] N. Wang, D. Tao, X. Gao, X. Li, and J. Li. A comprehensive survey to face hallucination. *IJCV*, 106(1):9–30, 2014.

[29] X. Wang and X. Tang. Hallucinating face by eigentransformation. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 35(3):425–434, 2005.

[30] C.-Y. Yang, S. Liu, and M.-H. Yang. Structured face hallucination. In *CVPR*, pages 1099–1106, 2013.

[31] C.-Y. Yang and M.-H. Yang. Fast direct super-resolution by simple functions. In *ICCV*, pages 561–568, 2013.

[32] J. Yang, J. Wright, T. S. Huang, and Y. Ma. Image super-resolution via sparse representation. *IEEE Transactions on Image Processing*, 19(11):2861–2873, 2010.

[33] X. Yu and F. Porikli. Ultra-resolving face images by discriminative generative networks. In *ECCV*, pages 318–333, 2016.

[34] S. Zhao, H. Han, and S. Peng. Wavelet-domain HMT-based image super-resolution. In *ICIP*, volume 2, pages 953–956, 2003.

[35] S. Zhu, S. Liu, C. C. Loy, and X. Tang. Deep cascaded bi-network for face hallucination. In *ECCV*, pages 614–630, 2016.

[36] X. W. Ziwei Liu, Ping Luo and X. Tang. Deep learning face attributes in the wild. In *ICCV*, 2015.