International Journal of Computer Vision (2019) 127:763–784
https://doi.org/10.1007/s11263-019-01154-8

Wavelet Domain Generative Adversarial Network for Multi-scale Face Hallucination

Huaibo Huang[1,2,3,4] · **Ran He**[1,2,3,4] · **Zhenan Sun**[1,2,3,4] · **Tieniu Tan**[1,2,3,4]

Received: 9 February 2018 / Accepted: 29 January 2019 / Published online: 12 February 2019

Abstract
Most modern face hallucination methods resort to convolutional neural networks (CNN) to infer high-resolution (HR) face images. However, when dealing with very low-resolution (LR) images, these CNN based methods tend to produce over-smoothed outputs. To address this challenge, this paper proposes a wavelet-domain generative adversarial method that can ultra-resolve a very low-resolution (like 16×16 or even 8×8) face image to its larger version of multiple upscaling factors ($2\times$ to $16\times$) in a unified framework. Different from the most existing studies that hallucinate faces in image pixel domain, our method firstly learns to predict the wavelet information of HR face images from its corresponding LR inputs before image-level super-resolution. To capture both global topology information and local texture details of human faces, a flexible and extensible generative adversarial network is designed with three types of losses: (1) wavelet reconstruction loss aims to push wavelets closer with the ground-truth; (2) wavelet adversarial loss aims to generate realistic wavelets; (3) identity preserving loss aims to help identity information recovery. Extensive experiments demonstrate that the presented approach not only achieves more appealing results both quantitatively and qualitatively than state-of-the-art face hallucination methods, but also can significantly improve identification accuracy for low-resolution face images captured in the wild.

Keywords Face hallucination · Super-resolution · Wavelet transform · Generative adversarial network · Face recognition

1 Introduction

Face hallucination, also known as face super-resolution (SR), refers to generating high-resolution (HR) face images from their corresponding low-resolution (LR) inputs. It is significant for most face-related applications, e.g. face recognition (Shamir 2008; Hayat et al. 2017), where most captured faces in the wild are low-resolution and lacking in essential facial details. It is a domain-specific single image super-resolution (SISR) problem and many methods (Wang and Tang 2005; Liu et al. 2007; Yang et al. 2008; Park and Lee 2008; Li et al. 2009; Ma et al. 2010; Jung et al. 2011; Yang et al. 2013; Wang et al. 2014; Jiang et al. 2014; Zhu et al. 2016; Yu and Porikli 2016, 2017a, b; Farrugia and Guillemot 2017; Yang et al. 2017) 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.

Recently, deep learning based methods have been introduced into single image super-resolution problem and made great improvements. However, most of these CNN (Convolutional Neural Networks) based methods (Bruna et al. 2016;

Communicated by Xiaoou Tang.

✉ Ran He
rhe@nlpr.ia.ac.cn

Huaibo Huang
huaibo.huang@cripac.ia.ac.cn

Zhenan Sun
znsun@nlpr.ia.ac.cn

Tieniu Tan
tnt@nlpr.ia.ac.cn

[1] School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing, China

[2] Center for Research on Intelligent Perception and Computing, CASIA, Beijing, China

[3] National Laboratory of Pattern Recognition, CASIA, Beijing, China

[4] Center for Excellence in Brain Science and Intelligence Technology, CAS, Beijing, China

Dong et al. 2016; Kim et al. 2016a, b; Shi et al. 2016; Tai et al. 2017; Lai et al. 2017) depend on a pixel-wise mean squared error (MSE) loss in image pixel domain to push the outputs pixel-wise closer to the ground-truth HR images in training phase. Such approaches ignore the conditional dependency between super-resolved pixels and thus tend to produce blurry and over-smoothed outputs, lacking high-frequency textural details. Besides, they seem to only work well on limited up-scaling factors (2× to 4×) and degrade greatly when ultra-resolving a very small input (like 16×16 or smaller). Specifically for face hallucination, several recent efforts (Dahl et al. 2017; Yu and Porikli 2016, 2017a, b; Zhu et al. 2016) have been made to deal with ultra-resolving very small faces based on convolutional neural networks. Dahl et al. (2017) use PixelCNN (van den Oord et al. 2016) to synthesize realistic details. Yu and Porikli (2016, 2017a, b) investigate GAN (Goodfellow et al. 2014) to create perceptually realistic results. Zhu et al. (2016) combine dense correspondence field estimation with face hallucination. However, the applications of these methods for the super-resolution in image pixel domain face many problems, such as computational complexity in sampling (Dahl et al. 2017), instability in training (Yu and Porikli 2016, 2017a, b), poor robustness for pose and occlusion variations (Zhu et al. 2016). Moreover, the existing face hallucination methods mainly use visual perceptual results and standard image quality metrics such as PSNR and SSIM to evaluate their performance, which is inadequate to demonstrate whether the recovered information is helpful for face-related applications, e.g. face recognition. Therefore, due to various problems yet to be solved, face hallucination 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 (Mallat 1996). It can depict the contextual and textural information of an image at different levels, which motivates us to introduce wavelet transform to a deep super-resolution system. As illustrated in Fig. 1, the approximation coefficients, i.e. the top-left patches in Fig. 1b–d, of different-level wavelet packet decomposition (Coifman and Wickerhauser 1992) compress the face's global topology information at different levels; the detail coefficients, i.e. the rest patches in Fig. 1b–d, reveal face's structure and texture information. While the approximation coefficient can be seen as the down-sampled low-resolution version of a high-resolution image, super-resolution can be approximately considered as the inverse process of wavelet decomposition with the inferred detail coefficients. We assume that a high-quality high-resolution image with abundant textural details and invariant global topology information can be reconstructed via a low-resolution image as long as the corresponding wavelet coefficients are accurately predicted. Hence, the task of inferring a high-resolution face

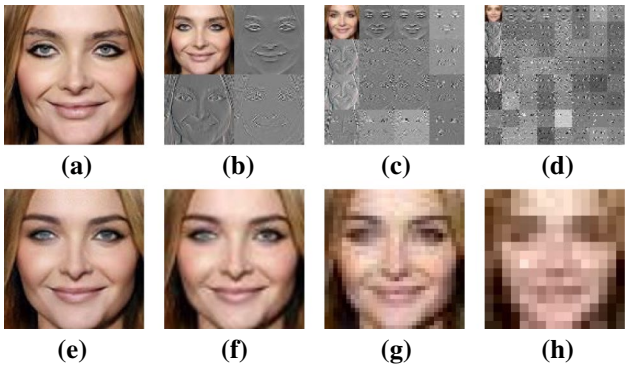


Fig. 1 Illustration of wavelet decomposition and our wavelet-domain face hallucination. Top row: **a** the original 128×128 HR face image and its **b** 1 level, **c** 2 level, **d** 3 level, full wavelet packet decomposition image. Bottom row: **h** the 16×16 low-resolution face image and its **g** 2×, **f** 4×, **e** 8×, upscaling versions inferred by our approach. While the approximation coefficient can be seen as the down-sampled low-resolution version of a high-resolution face, face hallucination can be approximately considered as the inverse process of wavelet decomposition with the inferred detail coefficients

is transformed to predicting a series of wavelet coefficients. Emphasis on the prediction of high-frequency wavelet coefficients helps recover texture details, while constraints on the reconstruction of low-frequency wavelet coefficients enforce consistence on global topology information. The combination of the two aspects makes the final high-resolution results more photo-realistic.

To take full advantage of wavelet transform, we present a wavelet-domain generative adversarial network (WaveletSRGAN) for face hallucination with three types of losses: wavelet reconstruction loss to push wavelets closer with the ground-truth, wavelet adversarial loss to generate perceptually realistic wavelets, and identity preserving loss to help identity information recovery. Coordinated with these losses, as outlined in Fig. 2, WaveletSRGAN contains three parts: wavelet-domain super-resolution network (WaveletSRCNN), wavelet-domain discriminator network (WaveletDNet) and facial evaluation network (EvalNet).

The wavelet-domain super-resolution network (WaveletSRCNN) takes a low-resolution face as an input and predicts the corresponding series of wavelet coefficients before reconstructing the high-resolution outputs. It can be further decomposed into three subnetworks: embedding, wavelet prediction and reconstruction networks. The embedding net represents a low-resolution face image to a set of feature maps before up-scaling. 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 upscaling factor flexible as well. Besides, as each wavelet coefficient shares the same size with the low-resolution input, the network

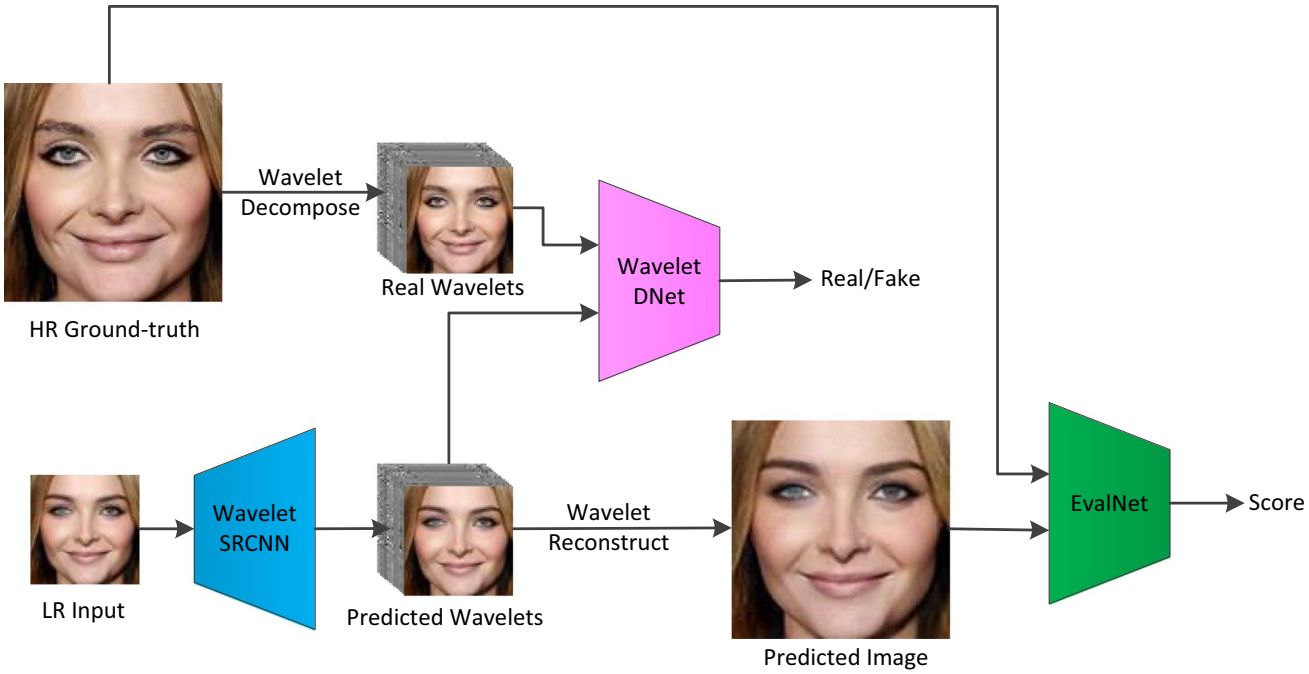


Fig. 2 The proposed wavelet-domain generative adversarial network (WaveletSRGAN) for face hallucination comprises wavelet-domain super-resolution network (WaveletSRCNN), wavelet-domain discriminator network (WaveletDNet) and facial evaluation network (EvalNet). WaveletSRCNN is a CNN that takes a low-resolution face image as an input, predicts the corresponding wavelets and outputs the desired high-resolution image. WaveletDNet is trained adversarially against WaveletSRCNN to distinguish the generated and real wavelets. EvalNet is adopted to evaluate and improve the recovered identity information via face hallucination

configuration is selected to keep every feature map the same size with the input, which reduces the difficulty of training. The reconstruction network is used to reconstruct the expected HR image from the inferred wavelet coefficients, acting as a fixed learned matrix. As the proposed 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.

The wavelet-domain discriminator network (WaveletDNet) is adopted to distinguish fake and real wavelets rather than fake and real images, which is different from other GAN based SR methods (Sønderby et al. 2017; Ledig et al. 2017; Yu and Porikli 2016, 2017a, b; Sajjadi et al. 2017; Xu et al. 2017). Through adversarial training, the proposed WaveletSRCNN tends to produce realistic wavelets, namely realistic low-frequency wavelets for global facial topology information and realistic high-frequency wavelets for local facial texture details. WaveletDNet consists of three parts: the wavelet embedding network, the sum operator and the prediction network. The wavelet embedding network employs multiple independent subnets to map each wavelet into an individual set of feature maps. These feature maps are fused using sum operation and then fed to the prediction net, which makes the discriminator network also flexible like WaveletSRCNN. Besides, as the wavelets have the same small size with the low-resolution inputs, WaveletDNet can be designed with a very shallow architecture to ensure enough receptive field, which reduces the difficulty in adversarial training for high-resolution images.

The facial evaluation network (EvalNet) is designed to evaluate and improve the recovered useful facial information, i.e. identity information in this paper, via face hallucination. It takes the hallucinated and ground-truth faces as inputs and compute the similarity scores to evaluate how much useful facial information has been recovered. For the reason that identity is the most important intrinsic facial information, we select a pretrained face recognition model as EvalNet and propose an identity preserving loss to help the recovery of identity information. Moreover, face verification metrics are employed to evaluate the identity recovery performance of face hallucination.

The main contributions are summarized as follows:

- A novel wavelet-domain approach is proposed for deep face hallucination. To the best of our knowledge, this is the first attempt to transform single image super-resolution to wavelet coefficients prediction task in deep learning framework - albeit many wavelet-domain researches exist for super-resolution.
- A flexible and extensible fully convolutional neural network is presented to make the best use of wavelet trans-

form. It can apply to faces of different input-resolutions with multiple upscaling factors.
- A simple yet effective wavelet-domain discriminator network is proposed for face hallucination. It simplifies the complexity of discriminator architecture and reduces the difficulty in training GAN for high-resolution images.
- A facial evaluation network with identity preserving loss is proposed to improve the recovery of identity information. Face verification metrics are also introduced to face hallucination.
- Experimental results on multi-scale face hallucination, especially on very small faces, show that the proposed approach outperforms state-of-the-art methods in terms of both traditional super-resolution metrics and face verification metrics.

This paper is an extension of our previous conference version (Huang et al. 2017). Apart from providing more in-depth analysis and more extensive experiments, the major difference between this paper and its previous version lies in three-folds: 1) Face hallucination is extended from wavelet-domain convolutional neural network (WaveletSRCNN) to wavelet-domain generative adversarial network (WaveletSR-GAN) with the additions of a wavelet-domain discriminator and a facial evaluation net. Two new loss functions are used to synthesize photo-realistic textures and preserve identity information. 2) A simple yet effective wavelet-domain discriminator is proposed to generate perceptually plausible wavelets. It simplifies the discriminator architecture and reduces the training difficulty for GAN to generate high-resolution images. 3) Identity verification evaluation is introduced to face hallucination and demonstrate the advantage of the proposed method compared with the state-of-the-arts. Our new adversarial method further improves the verification rate on the LFW (Huang et al. 2007) database from 68.33% to 81.20% (false acceptance rate is at 0.1%) when the probe faces are of low-resolution 16×16 pixel-size before being hallucinated.

2 Related Work

Face hallucination is a specific case of single image super-resolution, which is extended to wavelet-domain generative adversarial network in this paper. In this section, we briefly review some related advances in generic single image super-resolution, face hallucination, wavelet-domain super-resolution and generative adversarial network.

2.1 Single Image Super-Resolution

In general, single image super-resolution methods can be divided into three types: interpolation-based, statistics-based and learning-based methods. In the early years, the former two types (Sun et al. 2008; Yang et al. 2010; Yang and Yang 2013) have attracted most of attention for their computationally efficiency. However, they are always limited to small upscaling factors. Learning based methods (Chang et al. 2004; Lin et al. 2008; Singh et al. 2014; Huang et al. 2015) employ large quantities of LR/HR image pair data to infer missing high-frequency information and promise to break the limitations of big magnification.

Recently, deep learning based methods (Dong et al. 2016; Kim et al. 2016a, b; Shi et al. 2016; Lai et al. 2017; Tai et al. 2017; Tong et al. 2017) have been introduced into super-resolution problem due to their powerful ability to learn knowledge from large databases. Dong et al. (2016) incorporate convolutional neural networks to directly learn an end-to-end mapping between the low/high-resolution images. The following researchers explore various methods to improve CNN-based super-resolution through deeper and more complex networks. However, most of these convolutional methods depend on MSE loss to learn a map function of LR/HR image pairs, which leads to over-smoothed outputs when the input resolution is very low and the magnification is large.

Several works have been recently presented to alleviate this problem. To improve the outputs' perceptual quality, Johnson et al. (2016) and Bruna et al. (2016) propose the perceptual loss based on the high-level features extracted from pretrained networks. Sajjadi et al. (2017) propose the texture matching loss to create realistic textures. Ledig et al. (2017) propose a generative adversarial network (SRGAN) for image super-resolution, which optimizes a combination function of an adversarial loss and a content loss. Though the architecture of our network is similar with theirs to some degree, we hallucinate face images in wavelet domain rather than image pixel domain; our version of the perceptual loss is facial specific to help identity preserving; our method works well on very small faces with large upscaling factors like 8× to 16× while SRGAN focuses on 4×.

2.2 Face Hallucination

Specific to face hallucination, many methods (Wang and Tang 2005; Jung et al. 2011; Wang et al. 2014; Yang et al. 2013; Zhu et al. 2016) are proposed to exploit the specific static information of face images with the help of face analysis technique. Wang and Tang (2005) estimate landmarks and facial pose before reconstructing high-resolution images while the accurate estimation is difficult for rather small faces. Zhu et al. (2016) present a unified framework of face hallucination and dense correspondence field estimation to recover textural details. They achieve appealing results for low-resolution faces but cannot work well on faces with various poses and occlusions, due to the difficulty of accurate spatial prediction.

Similarly with single image super-resolution, generative models (Yu and Porikli 2016, 2017a, b; Dahl et al. 2017) are also brought to face hallucination to learn face prior knowledge. Yu and Porikli (2016) propose a generative adversarial network to resolve 16×16 pixel-size faces to its 8× larger versions. Dahl et al. (2017) present a recursive framework based on PixelCNN (van den Oord et al. 2016) to synthesize details of 4× magnified images with 8×8 low-resolution inputs. The 32×32 outputs are not sufficiently perceptual appealing, and their method suffers from high computational complexity when sampling high-resolution images.

Recently, several new approaches have been proposed to tackle with similar tasks. Most of these approaches try to explore prior information or attributes to facilitate face hallucination. Bulat and Tzimiropoulos (2018) localize the facial landmarks on the hallucinated faces and improve super-resolution through a heatmap loss. Chen et al. (2018) estimate facial prior on the coarse super-resolved faces to help the fine super-resolution subnet. Yu et al. (2018a) present a multi-task framework to exploit image intensity similarity and explore the face structure prior simultaneously. Yu et al. (2018b) utilize supplemental attributes to reduce the ambiguity in face hallucination. Moreover, Bulat et al. (2018) train a high-to-low GAN to simulate the image degradation process and another low-to-high GAN for real-world super-resolution. On contrast to these methods, our wavelet domain method provides a basic super-resolution architecture. We can easily combine these priors into our method or apply it to other related problems.

2.3 Wavelet-Domain Super-Resolution

Many wavelet-domain methods have already been proposed for super-resolution problem. A large percentage of them (Nguyen and Milanfar 2000; Ji and Fermüller 2009) focus on multiple images super-resolution, 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 (Anbarjafari and Demirel 2010; Naik and Patel 2013) and statistic-based (Zhao et al. 2003) methods. Naik et al. (Naik and Patel 2013) propose a modified version of classical wavelet-domain interpolation method (Anbarjafari and Demirel 2010). Gao et al. (2016) propose a hybrid wavelet convolution network. They use wavelet to provide a set of sparse coding candidates and utilize another convolution net for sparse coding, which is totally different from ours. Besides, Mallat (2016) uses wavelet transform to separate the variations of data at different scales, while we predict the wavelets from low-resolution inputs for face hallucination.

2.4 Generative Adversarial Networks

Our work is also related to the generative adversarial networks(GAN) (Goodfellow et al. 2014), which trains generator and discriminator via a min-max two player game to learn image prior. It is easy for discriminator to distinguish the generated and real high-resolution images, which harms the balance of adversarial training and makes it difficult for GAN to generate realistic high-resolution images. While many efforts (Zhang et al. 2017; Karras et al. 2018; Huang et al. 2018) have been made to address this problem, our proposed wavelet adversarial method provides another simple yet efficient way to reduce the training difficulty in generating high-resolution images via low-resolution wavelets.

3 Approach

In this section, we present a novel wavelet-domain framework for face hallucination, which predicts a series of corresponding wavelet coefficients instead of high-resolution images directly. Three types of losses are proposed to generate realistic wavelets and recover identity information. Then, a flexible and extensible wavelet-domain generative adversarial network (WaveletSRGAN) is designed for multi-scale face hallucination. At last, the implementation details of WaveletSRGAN are given.

3.1 Wavelet Transform

In order to further illustrate the motivation of hallucinating faces in wavelet domain, we first explore the relationship between the high-frequency wavelets and the image blur level. Given a high-resolution face image of 128×128 pixel-size, a series of blurry images are synthesized through down-sampling following by up-sampling using bicubic interpolation. The responding wavelets are achieved by 2-level haar wavelet packet decomposition. We sample 10K images of each blur level and then compute the mean absolute value (AVG_HF) of the detail wavelet coefficients. As shown in Fig. 3, both the visual and quantitative results demonstrate that high-frequency wavelets fade along with the increase of the blur level. In other words, it is essential to recover high-frequency wavelets as more as possible for generating shaper images. As described in the first section, deep super-resolution networks with the reconstruction loss in image domain tend to generate blurry images, which illustrates that those methods fail to predict high-resolution images with abundant high-frequency wavelets. To alleviate this problem, we resort to hallucinating faces directly in

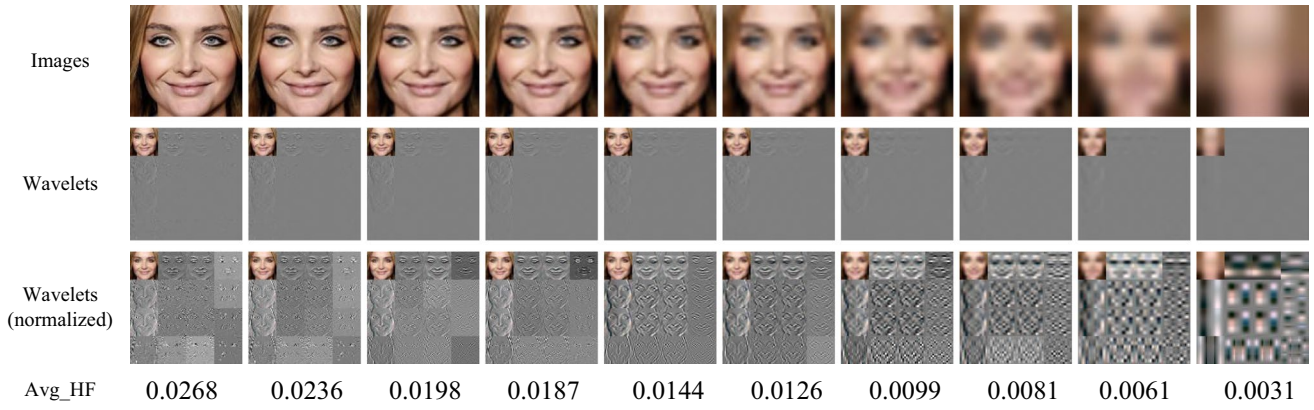


Fig. 3 The relationship between the high-frequency wavelets and the image blur level. The top row are the images of different blur level (which are generated from low-resolution images using bicubic interpolation; the input resolutions are 128, 96, 64, 48, 32, 24, 16, 12, 8, and 4, respectively). The second and third rows are the responding wavelets and the responding normalized wavelets for better view. The bottom row are the mean absolute values of the high-frequency wavelets (i.e., all the detail wavelet coefficients)

wavelet domain for recovering high-frequency details while preserving global facial information.

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) (Mallat 1989) to compute Haar wavelets. The wavelet coefficients at different levels are computed by repeating the decomposition in Fig. 4 to each output coefficient iteratively. Example results of wavelet packet transform at different levels are showed in Fig. 1b–d.

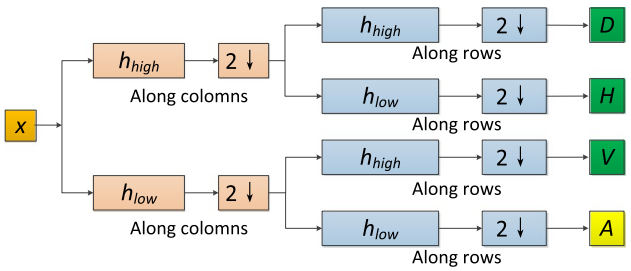


Fig. 4 Illustration of fast wavelet transform (FWT). FWT uses low-pass and high-pass decomposition filters iteratively to compute wavelet coefficients, where Haar-based $h_{low} = (1/\sqrt{2}, 1/\sqrt{2})$ and $h_{high} = (1/\sqrt{2}, -1/\sqrt{2})$

3.2 Loss Function

3.2.1 Wavelet Reconstruction Loss

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

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

The most common loss function is pixel-wise MSE in image pixel domain, defined as

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

As argued in many papers (Ledig et al. 2017; Yu and Porikli 2016; Sønderby et al. 2017; Dahl et al. 2017), 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 the super-resolution problem from the original image pixel domain to the wavelet domain and introduce wavelet-domain loss functions to help texture reconstruction.

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

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

We propose a weighted version of MSE loss in wavelet domain to optimize the above object, defined as

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

where $W = (\lambda_1, \lambda_2, \ldots, \lambda_{N_w})$ 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 an input.

3.2.2 Wavelet Adversarial Loss

In addition to the wavelet reconstruction loss, we also employ a wavelet adversarial loss to encourage the reconstructed wavelets to obey the prior distribution of realistic facial wavelets. In the original setting (Goodfellow et al. 2014), generator network G is trained to learn a mapping from noise variables to a data space of images; discriminator network D is trained to distinguish between real and generated images. In our work, we use the proposed WaveletSRCNN as the generator, which outputs the wavelets conditioned on the low-resolution inputs. The discriminator network is designed to distinguish between real and generated wavelets.

We adopt the Least Squares Generative Adversarial Networks (LSGANs) (Mao et al. 2017) to train our network. The wavelet adversarial losses for generator and discriminator are defined as

$$
l_{wavelet\_adv} = \frac{1}{2L} \|D(G(x)) - 1\|_F^2, \tag{5}
$$

$$
l_{wavelet\_dis} = \frac{1}{2L} \|D(C) - 1\|_F^2 + \frac{1}{2L} \|D(G(x)) - 0\|_F^2, \tag{6}
$$

where x is the low-resolution input, C contains the ground-truth wavelets, and L is the size of the output of the discriminator.

3.2.3 Identity Preserving Loss

An identity preserving loss is proposed to recover useful facial information. It is motivated by the ideas of Gatys et al. (2016), Johnson et al. (2016) and Bruna et al. (2016), Sohn et al. (2017) that semantic information can be represented at different levels by the features extracted from pretrained networks. Perceptual losses defined on high-level features can improve image perceptual quality. As our main purpose is to enforce the hallucinated faces to have small distances with the ground-truths in a facial semantic space, we select a pretrained face recognition network, i.e. LightCNN (Wu et al. 2018), to extract high-level features.

The identity preserving loss based on LightCNN is defined as

$$
\begin{aligned}
l_{identity} &= \sum_{i=1}^{N_f} \frac{1}{L_i} \|F_i(\hat{y}) - F_i(y)\|_1 \\
&= \sum_{i=1}^{N_f} \frac{1}{L_i} \|F_i(R\hat{C}) - F_i(y)\|_1,
\end{aligned} \tag{7}
$$

where F_i is the i-th layer of N_f feature extractor layers, L_i is the size of the i-th layer, y is the ground-truth face, $\hat{y} = R\hat{C}$ and $\hat{C}$ are the generated image and wavelets respectively, R is the reconstruction matrix to reconstruct HR images from wavelets.

3.2.4 Overall Loss Function

To sum up, the full objective for generator WaveletSRCNN is a weighted sum of all the losses defined above: $l_{wavelet\_rec}$ to force wavelet reconstruction, $l_{wavelet\_adv}$ to generate realistic wavelets, $l_{identity}$ to recover identity information.

$$
l_G = \alpha_1 l_{wavelet\_rec} + \alpha_2 l_{wavelet\_adv} + \alpha_3 l_{identity} \tag{8}
$$

where α_1, α_2 and α_3 are weighting coefficients to balance each item.

3.3 Network Architecture

As outlined in Fig. 2, our proposed network for face hallucination (WaveletSRGAN) consists of three subnetworks: wavelet-domain convolutional neural network (WaveletSRCNN), wavelet-domain discriminator network (WaveletDNet) and facial evaluation network (EvalNet). In the following, the architectures of each subnetwork are described in detail.

3.3.1 Architecture of WaveletSRCNN

As showed in Fig. 5, our wavelet-domain SR network consists of three subnetworks: embedding, wavelet prediction, and reconstruction networks. The embedding net represents a 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.

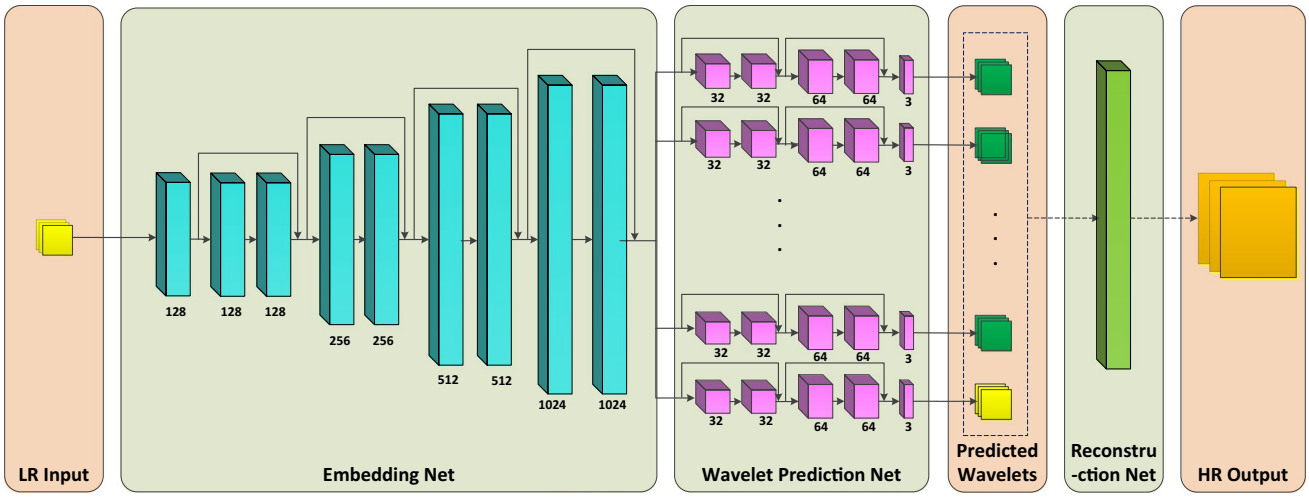


Fig. 5 The architecture of our wavelet-domain super-resolution net (Wavelet-SRNet). All the convolution layers have the same filter kernel-size of 3×3 and each number below them defines their individual output channel-size. Skip connections exist between every two convolution layers (except the first layer) in the embedding and wavelet prediction nets

The **embedding net** takes a low-resolution image of the size $3 \times h \times w$ as an input and represents it as a set of feature maps. All the convolution filters share the same size of 3×3 with a stride of 1 and a pad of 1, which makes each 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 an input and generates the corresponding wavelet coefficients. We set all the convolution filters the size of 3×3 with a stride of 1 and a pad of 1 just like 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× and 8× 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 filter-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 (i.e., R in Eq. 7, the pre-computed parameters can be downloaded with the released code)[1] and fixed in training. Hence it has no effect on the extensibility of the whole networks. It is noted that the reconstruction net acts like a function to implement the wavelet reconstruction and can be done offline as post-processing step if the identity preserving loss is not used. It is designed as a deconvolution layer to allow the backward propagation of the gradients from the EvalNet and allow the end-to-end training of the whole network of WaveletSRGAN.

As mentioned above, all the convolution filters 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 each 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-normalization (Ioffe and Szegedy 2015) and Rectified Linear Unit (ReLU) are also used after each layer, except the last layer of the wavelet prediction net.

The definition of WaveletSRCNN can be formulated as follows

$$\begin{aligned} \hat{y} &= \phi(\hat{C}) = \phi\{(\hat{c}_1, \hat{c}_2, \ldots, \hat{c}_{N_w})\} \\ &= \phi\{(\varphi_1(\hat{z}), \varphi_2(\hat{z}), \ldots, \varphi_{N_w}(\hat{z}))\} \\ &= \phi\{(\varphi_1(\psi(x)), \varphi_2(\psi(x)), \ldots, \varphi_{N_w}(\psi(x)))\}, \end{aligned} \tag{9}$$

where

[1] https://github.com/hhb072/WaveletSRNet

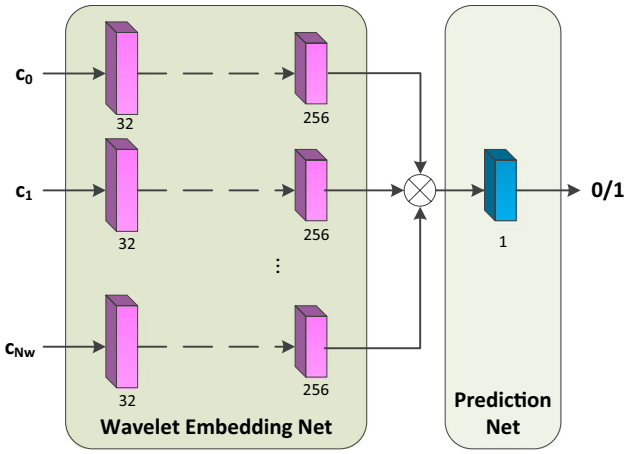


Fig. 6 The architecture of our wavelet-domain discriminator net (WaveletDNet). The wavelet embedding net consists of multiple individual subnets, each of which is a sequence of convolution layers with the kernel-size 3×3 following by batch-norm and leaky-relu. The stride-sizes are 2 except that of the last layer is 1. $\otimes$ means the sum operation. The prediction net is a single convolution layer with the kernel-size 3×3 and stride-size 1. Each number below the blocks is their output channel-size

$$\begin{aligned} &\psi : R^{3\times h\times w} \rightarrow R^{N_e\times h\times w}, \\ &\varphi_i : R^{N_e\times h\times w} \rightarrow R^{3\times h\times w}, \quad i = 1, 2, \ldots, N_w, \\ &\phi : R^{N_w\times 3\times h\times w} \rightarrow R^{3\times (r\times h)\times (r\times w)}, \end{aligned} \tag{10}$$

are mappings of the embedding, wavelet prediction, reconstruction nets, respectively. It is noted that N_e is fixed while N_w changes in accordance with the upscaling factor r, i.e., $N_w = r^2$.

3.3.2 Architecture of WaveletDNet

As showed in Fig. 6, our wavelet-domain discriminator net (WaveletDNet) consists of three parts: wavelet embedding network, sum operator, and prediction network. The **wavelet embedding network** takes the generated or real wavelets as an input and represents each wavelet into corresponding individual feature maps. It comprises N_w parallel independent subnets, where $N_w = 4^n$ on the condition of n level of wavelet-packet decomposition like the above wavelet prediction net in WaveletSRCNN. Each subnet takes a wavelet image of the size $3 \times h \times w$ as an input and represents it as a set of feature maps. Then all these sets of the feature maps are fused using **sum operation**. The **prediction net** map the fused features into a single channel of feature map, which is regressed to zero or one in the adversarial training. All the convolution layers in WaveletDNet have the same small kernel-size 3×3 and pad-size 1. The stride-sizes are 2 to reduce the size of feature maps, except those of the last layer in the wavelet embedding net and the prediction layer are 1.

As we use sum operation to fuse the output features, the number of the subnets in the wavelet embedding net is also flexible to change, just like that in WaveletSRCNN, which makes WaveletDNet is also extensible according to the upscaling factor. Besides, as the input wavelet has the same low-resolution with the LR input, a very shallow architecture is able to ensure enough receptive field. We use two convolution layers in the wavelet embedding net in this paper, the output channel-sizes of which are 32 and 256 respectively.

The definition of WaveletDNet can be formulated as follows

$$D(C) = W_p^T \sum_i^{N_w} \eta_i(c_i), \tag{11}$$

where W_p is the corresponding matrix of the convolution layer in the prediction net, η_i is the corresponding map function of the i-th subnet in the wavelet embedding net.

3.3.3 Architecture of EvalNet

We select the pretrained LightCNN (Wu et al. 2018) as our EvalNet. LightCNN is a face recognition network, which includes 29 convolution layers, 4 max-pooling layers, and one fully-connected layer. It is pre-trained to classify tens of thousands of identities, which make it has powerful ability to capture the most distinguishable feature for face identity information. We use the outputs of the last two layers of LightCNN to compute the similarity score and the loss function in Eq. (8). The parameters in LightCNN are fixed during training.

3.4 Implementation Details

A novel training trick for face hallucination, called as co-training, is used to make our model stable in training. Two types of low-resolution images are taken as the 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. All the face images are normalized with two eyes aligned horizontally and then center-cropped to 128×128 size, following Wu et al. (2018). Wavelet packet decomposition at 3 level is used to get the ground-truth wavelet coefficients c_i in Eq. (4). The approximation coefficient c_1 is treated as one version of the 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 the low-resolution input, directly down-sampled by bicubic interpolation, is used cooperatively with the wavelet version, which helps maintain the robustness of our model. In the testing phase, we evaluate on faces down-sampled by bicubic interpolation.

Since our generator network WaveletSRCNN is fully convolutional 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× magnification, and then fine-tune it for 8×8 input resolution with 8× magnification. For 8×8 input resolution with 16× magnification, we initialize the parameters by the overlapping ones of the model for 8×8 with 8× magnification before fine-tuning it. For other cases, we just choose the closest model for evaluation. Besides, as our discriminator network WaveletDNet is also fully convolutional, similar training tricks are taken for different upscaling factors.

We set the hyper-parameters empirically to balance the importance of different losses. The trade-off parameter α_1 for wavelet reconstruction loss is set to 1, α_2 for wavelet adversarial loss is set to 10, and α_3 for identity preserving loss is set to 10, λ_1 for the reconstruction of approximation wavelets is set to 0.01, $\lambda_2 \sim \lambda_{N_w}$ for the reconstruction of high-frequency wavelets are set to 0.99. We train generator WaveletSRCNN and discriminator WaveletDNet adversarially using Adam algorithm (Kingma and Ba 2014) with a batch size of 64 and a fixed learning rate of 0.0002. It takes about 30 epochs for our network to converge, among which the first 10 epochs are pretrained using only wavelet reconstruction loss.

4 Experiments

In this section, we evaluate the proposed approach against state-of-the-art generic super-resolution and face hallucination methods for multiple input-resolutions on two widely used face databases. Both qualitative results and quantitative results are reported, not only on traditional SR metrics (PSNR and SSIM) but also on face verification metrics. Ablation experiments are also conducted to demonstrate the effectiveness of each part of our model. The time complexity and robustness toward several hard cases is also discussed.

4.1 Datasets and Protocols

The CelebA database (Liu et al. 2015) The CelebFaces Attributes (CelebA) dataset is the mostly used super-resolution and face hallucination database. There are more than 200k celebrity face images that cover large pose and occlusion variations. Following the standard protocol, we divide the database into three subsets: the training set of 162,700 images, the validation set of 19,867 images and the testing set of 19,962 images. We use the training set to train our model and the validation set for validation during the training phase. We use the testing set to evaluate the SR performance in the testing phase. The standard image quality measures, i.e., PSNR and SSIM, are adopted as the quantitative metrics for super-resolution, where PSNR is calculated on the luminance channel, following Zhu et al. (2016) and SSIM is calculated on the three channels of RGB (It is noted that we use Matlab to calculate the two metrics, which may cause different values with those using other tools).

The LFW database (Huang et al. 2007) The Labeled Faces in the Wild (LFW) dataset is the commonly used database for unconstrained face recognition. There are 13,233 face images of 5,749 people captured in the wild. The dataset provides a standard protocol for face verification, which contains 6,000 face pairs with 3,000 positive pairs and 3,000 negative pairs. On one hand, we use the whole dataset to evaluate the SR performance using PSNR and SSIM like the CelebA testing set. On the other hand, we use the standard 6,000 face pairs to evaluate face verification performance via face hallucination. The images in the gallery set are kept unchanged and the images in the probe set are down-sampled to generate low-resolution samples. We hallucinate the down-sampled probe faces using our method and the state-of-the-arts and then compare the verification performance.

To sum up, we generate HR/LR sample pairs using the images in the CelebA and LFW database. The proposed network is trained on the CelebA training set and evaluated on the CelebA testing set and the whole LFW dataset, assuring no over-lapped images appearing in both the training and testing phase. We use two types of quantitative metrics for evaluation: PSNR and SSIM to evaluate traditional SR performance; face verification metrics to evaluate how much identity information has been recovered.

4.2 Results on Multiple Input Resolutions

As mentioned above, our method can apply to different input resolutions with multiple magnifications. In Fig. 7, we provide the qualitative results of our method on three input resolutions (8×8, 16×16 and 32×32) compared with the bicubic interpolation baseline.

As for the faces of 16×16 and 32×32 pixel-sizes, the hallucination results of our method both have a realistic looking and keep almost pixel-wise similarity with the ground-truth. As for the faces of 8×8 resolutions, the hallucination results are unable to keep accurately pixel-wise similarity, while still looking photo-realistic, which is different with the somewhat blurry results in our prior work (Huang et al. 2017). It is noted that facial attributes, like sketch, hair style, skin color and so on, can be excellently recovered for

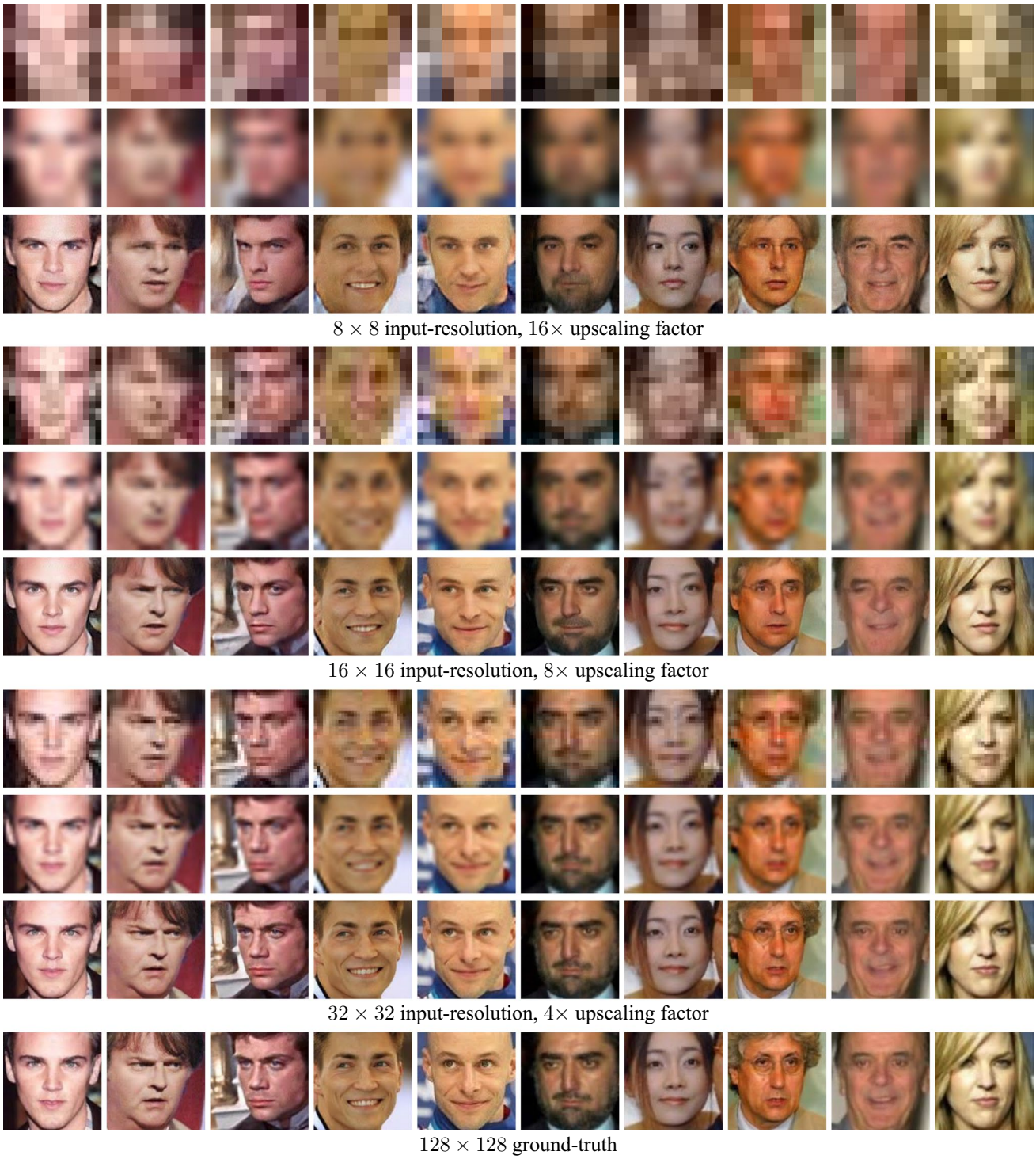


Fig. 7 Qualitative results of various input resolutions: 8×8, 16×16 and 32×32. For each input resolution, the first row is the low-resolution input faces, the second is the results of bicubic interpolation, and the third is ours. The bottom row is the ground-truth. From left to right, the former five and the latter five columns are randomly selected from the CelebA testing set and the LFW dataset, respectively. Best viewed by zooming in the electronic version

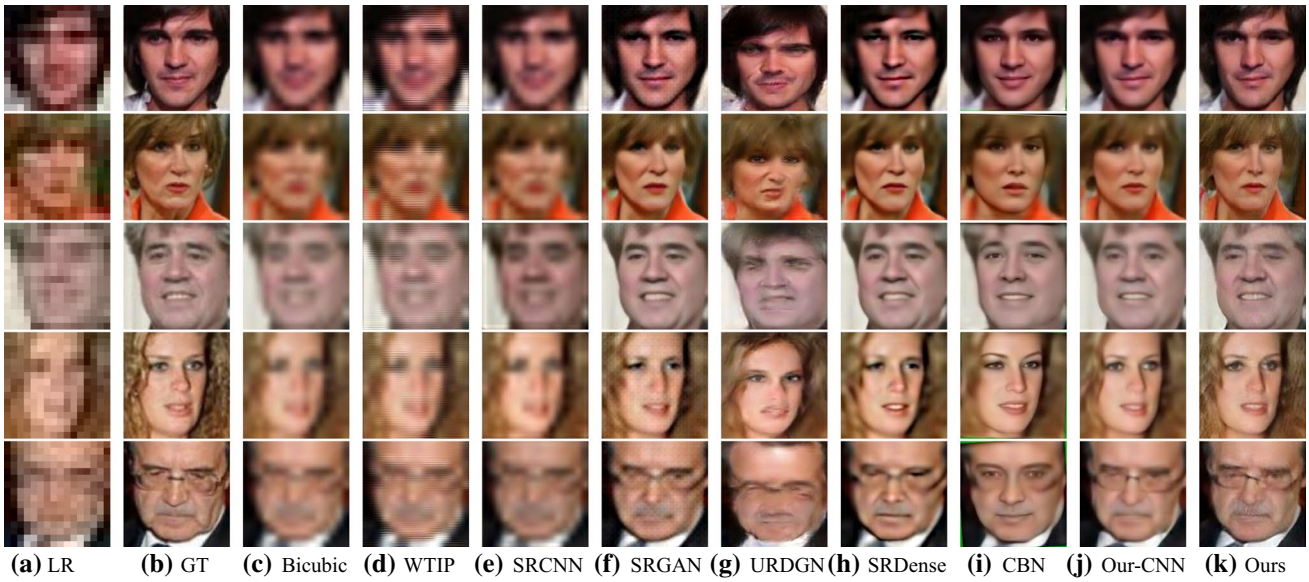


Fig. 8 Comparison with the state-of-the-art methods on 16×16 input resolution with 8× upscaling factor. All the images are randomly selected from the LFW dataset

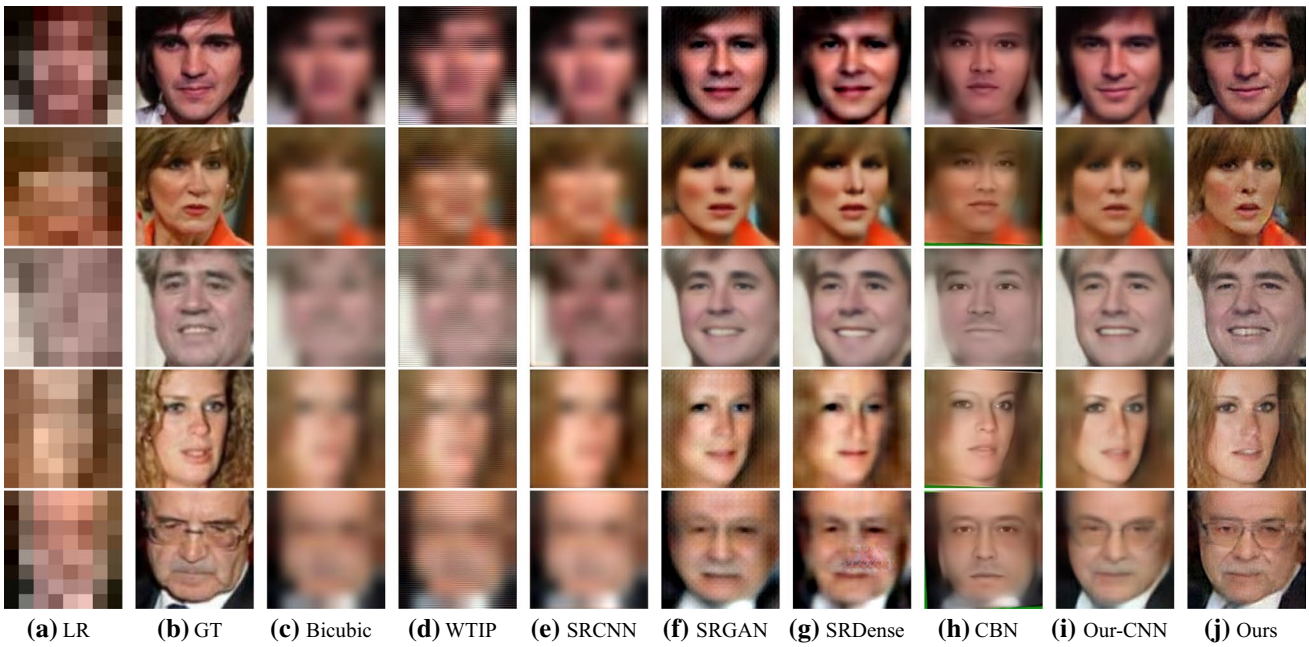


Fig. 9 Comparison with the state-of-the-art methods on 8×8 input resolution with 16× upscaling factor. All the images are randomly selected from the LFW dataset

all the three types of input resolutions. The inferred outputs are also perceptually identity-persistent to some degree. These facts imply that our method can recover abundant useful facial information from low-resolution faces, even for those very small ones of only 64 pixels.

4.3 Comparison on Very Low-Resolutions

We evaluate our method qualitatively on two very low-resolutions, i.e., 16×16 and 8×8, with the comparison with bicubic interpolation, wavelet-domain interpolation (WTIP for short) (Naik and Patel 2013) and state-of-the-art

Table 1 Quantitative comparison results on the CelebA testing set and the LFW dataset

Dataset	Settings	Metric	Bicubic	WTIP	SRCNN	SRGAN	URDGN	SRDense	CBN	Our-CNN	Ours
CelebA	$32 \times 32, 4\times$	PSNR	29.39	27.09	26.26	30.60	–	30.22	26.63	**33.81**	31.86
		SSIM	0.9320	0.8919	0.9092	0.9433	–	0.9386	0.8889	**0.9630**	0.9480
	$16 \times 16, 8\times$	PSNR	24.82	23.18	23.07	26.18	24.61	26.21	25.36	**28.15**	27.38
		SSIM	0.8599	0.7854	0.8425	0.8814	0.8525	0.8886	0.8810	**0.9107**	0.8916
	$8 \times 8, 16\times$	PSNR	21.43	20.52	20.18	22.67	–	21.55	21.61	**24.40**	23.18
		SSIM	0.7953	0.7041	0.7789	0.8067	–	0.7870	0.8076	**0.8507**	0.8141
LFW	$32 \times 32, 4\times$	PSNR	32.43	25.78	27.87	33.38	–	32.84	26.28	**37.07**	34.81
		SSIM	0.9628	0.8727	0.9442	0.9658	–	0.9639	0.9055	**0.9799**	0.9701
	$16 \times 16, 8\times$	PSNR	26.31	22.63	23.94	27.45	24.57	27.20	25.95	**29.44**	28.41
		SSIM	0.8948	0.7662	0.8773	0.9068	0.8604	0.9113	0.8986	**0.9317**	0.9117
	$8 \times 8, 16\times$	PSNR	22.28	19.86	20.61	22.98	–	22.15	22.24	**24.76**	23.40
		SSIM	0.8273	0.6849	0.8090	0.8249	–	0.8146	0.8300	**0.8675**	0.8250

Bold values indicate the best result

methods: SRCNN (Dong et al. 2016), SRGAN (Ledig et al. 2017), URDGN (Yu and Porikli 2016), SRDense (Tong et al. 2017), CBN (Zhu et al. 2016). We also compare the proposed method with our prior work (Huang et al. 2017), indicated as Our-CNN, which can be seen the generator WaveletSRCNN in Fig. 5 without wavelet adversarial loss and identity preserving loss. We retrain SRCNN, SRGAN and SRDense on the CelebA training set to suit better for face images. Since CBN uses the whole CelebA dataset for training, we give the qualitative results on LFW for a fair comparison, as showed in Figs. 8 and 9.

As for 16×16 input resolution in Fig. 8, our method achieves the best visual perceptual performance. As the input resolution is very low, there is very little information contained in the input images. Interpolation based methods like bicubic interpolation and WTIP cannot generate texture details and their results are rather over-smoothed. SRCNN is a very shallow CNN with three convolution layers and lack the ability to learn the map function between LR/HR pairs from such low input resolution. SRGAN and SRDenese have very deep architectures, which are enough to learn the map function. CBN uses a deep cascaded bi-network that works well when face landmarks can be accurately located. These above three methods achieve relatively high-quality results with the most facial information recovered. Our-CNN infers the corresponding wavelets from LR inputs to capture both the global topology information and local texture details and achieves comparable or a little better perceptual results than SRGAN, SRDense and CBN. The proposed method in this paper infers the most high-frequency details, like hair and beard, and its results look the most photo-realistic while keeping the pixel-similarity with the ground-truth at some degree.

As for 8×8 input resolution in Fig. 9, our method still achieves the best visual perceptual performance. This is a very hard case for face hallucination because only 64 pixels are contained in the input image. Bicubic interpolation, WTIP and SRCNN can only generate very over-smoothed facial contours. SRGAN and SRDense infer facial features but the results are blurry and closer to the mean face rather than the ground-truth. CBN generates mean-face like results for lack of the ability to accurately locate facial landmarks. Our-CNN produces better results in which global information is maintained and local details are recovered, while its results are also a little blurry on the edges and textures. Different from all the above methods, the proposed method can still hallucinate photo-realistic faces in this hard case. Even if the results of our method cannot keep accurately pixel-wise similar with the ground-truth, they seem to preserve identity information much better than others.

4.4 SR Quantitative Results

We evaluate our method quantitatively using standard SR quantitative measures, i.e., PSNR (db) and SSIM, on the CelebA testing set and LFW dataset. The evaluations are conducted in three cases: $(32 \times 32, 4\times)$, $(16 \times 16, 8\times)$ and $(8 \times 8, 16\times)$, where $(m \times n, r\times)$ means $m \times n$ input resolution with magnification factor r.

As shown in Table 1, Our-CNN and the proposed method achieve the best and the second best quantitative performances in the most cases. Our-CNN is a wavelet-domain convolutional neural network for face hallucination, which can be seen as a simplified version of the proposed method without the wavelet adversarial loss and identity preserving loss. The reason for it achieving the highest PSNR and SSIM values may be that the wavelet reconstruction loss in wavelet domain is helpful to minimize the reconstruction error in image pixel domain. The proposed method achieves a little lower quantitative values

Table 2 Comparison against Dahl et al. (2017) on CelebA faces of 8×8 pixels with magnification factor 4

Method	PSNR	SSIM	GPU-time (s)	CPU-time (s)
Bicubic	28.92	0.84	–	0.000514
Dahl et al. (2017)	29.09	0.86	53.01	382.4
Ours	**29.12**	**0.9121**	0.009205	0.3258

Bold values indicate the best result

than Our-CNN because it tends to pursue more perceptually plausible results rather than those merely with the minimization of MSE losses, as shown in Figs. 8 and 9. Even so, the proposed method outperforms the state-of-the-arts quantitatively, which implies that it not only provides perceptually plausible results with abundant texture details, but also preserves pixel-wise consistence with the ground-truth at some degree.

Besides, we also compare the proposed method with PixelCNN-based SR method (Dahl et al. 2017) on CelebA faces of 8×8 pixels with magnification factor 4. The results in Table 2 show that our method achieves better quantitative performance while runs much faster.

4.5 Face Verification Results

Besides the standard image quality measures like PSNR and SSIM, we introduce face verification metrics to evaluate the recovery of identity information via face hallucination. We conduct face verification experiments on the LFW dataset following the protocol described in the Sect. 4.1. The images in the probe set are down-sampled and then super-resolved. Face verification is conducted between the super-resolved probe set and the original gallery set. Two publicly released face recognition models are tested, i.e., the LightCNN (Wu et al. 2018) and the VGG-Face (Parkhi et al. 2015). The area under the ROC curve (AUC), true accept rates at 1% and 0.1% (TPR@FAR = 1%, TPR@FAR = 0.1%) are taken as evaluation metrics. The results in three settings of LR input resolutions and upscaling factors are reported in Table 3.

We use the hallucinated probe set by bicubic interpolation as baseline to demonstrate the influence of input resolutions on face verification. Take the metric TPR@FAR = 1% by LightCNN as example, it can be seen that the verification performance degrades greatly when the input resolution decreases, where the values are 97.77%, 96.10%, 45.50%, 3.17% for input-resolutions 128×128, 32×32, 16×16 and 8×8 pixel-size, respectively. This illustrates that face hallucination is important for low-resolution face recognition. We employ different face hallucination methods on the LR probe faces and find that most of them are helpful to face verification. As for 32×32 input-resolutions, the proposed method and Our-CNN outperform the others and achieve almost equivalent performance, where the values of TPR@FAR = 1% are 97.03% and 97.40%, respectively. As for the other two cases, 16×16 and 8×8 input-resolutions, the proposed method and Our-CNN achieve the best and the second

Table 3 Face verification results on the LFW dataset

Model	Settings	Metric	Original	Bicubic	WTIP	SRCNN	SRGAN	URDGN	SRDense	CBN	Our-CNN	Ours
LightCNN	32×32, 4×	AUC	99.31	99.16	99.04	99.17	99.22	–	99.21	90.80	99.25	**99.28**
		FAR = 1%	97.77	96.10	95.83	96.23	96.93	–	96.90	46.77	**97.40**	97.03
		FAR = 0.1%	96.23	91.90	91.70	92.87	94.07	–	94.97	32.53	95.73	**96.10**
	16×16, 8×	AUC	99.31	90.68	89.97	91.42	96.77	93.60	96.35	89.98	97.92	**98.48**
		FAR = 1%	97.77	45.50	40.53	48.70	78.83	53.57	77.50	46.90	87.97	**90.86**
		FAR = 0.1%	96.23	21.17	24.47	23.50	56.60	27.10	57.03	31.13	68.33	**81.20**
	8×8, 16×	AUC	99.31	60.89	59.40	61.47	77.10	–	74.30	63.00	87.29	**89.40**
		FAR= = 1%	97.77	3.17	2.90	2.83	16.40	–	12.67	4.57	38.43	**42.87**
		FAR= = 0.1%	96.23	0.27	0.47	0.30	4.23	–	3.73	1.30	12.93	**22.83**
VGG-Face	32×32, 4×	AUC	99.33	98.97	98.82	99.02	99.07	-	99.07	90.34	99.19	**99.21**
		FAR= = 1%	89.90	88.23	84.70	88.43	88.43	–	88.63	35.67	**89.40**	88.83
		FAR= = 0.1%	79.63	68.50	71.77	69.03	75.53	–	77.03	11.73	**78.27**	77.97
	16×16, 8×	AUC	99.33	86.61	82.85	88.56	96.08	91.52	95.64	89.54	97.50	**98.07**
		FAR= = 1%	89.90	24.77	19.53	29.47	62.70	44.37	56.93	34.33	73.77	**77.50**
		FAR = 0.1%	79.63	6.93	3.23	10.20	30.37	20.23	32.53	12.60	39.60	**55.87**
	8×8, 16×	AUC	99.33	54.78	53.92	56.86	80.10	–	77.48	65.67	89.92	**92.04**
		FAR = 1%	89.90	1.73	1.90	1.97	12.50	–	12.27	4.37	31.37	**41.33**
		FAR = 0.1%	79.63	0.30	0.13	0.17	3.67	–	2.13	0.50	3.63	**9.10**

Bold values indicate the best result

The images in the probe set are down-sampled and then super-resolved. We conduct face verification on the transformed probe set and the original gallery set. Results of 'Original' are obtained by directly testing on the original probe set

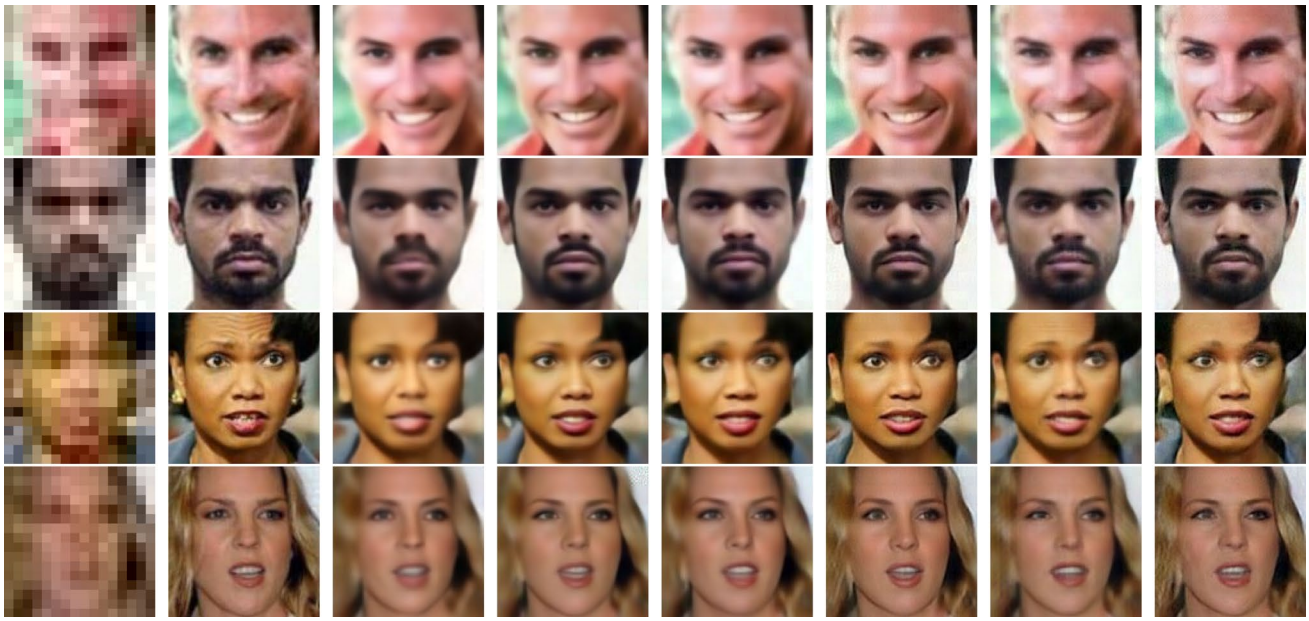


Fig. 10 Ablation: qualitative results of WaveletSRGAN and its variants on 16×16 input resolution with 8× upscaling factor. All the images are randomly selected from the LFW dataset. Best viewed by zooming in the electronic version

Table 4 Ablation: quantitative results of WaveletSRGAN and its variants on the LFW dataset with the input-resolution 16×16 and the upscaling factor 8×

Metric	ImgRec	ImgAdv	Our-CNN	Our-CNN+adv	Our-CNN+id	Ours
PSNR	29.04	29.07	**29.44**	28.79	29.42	28.41
SSIM	0.9215	0.9265	**0.9317**	0.9228	0.9306	0.9117
AUC	97.64	98.06	97.92	98.07	98.44	**98.48**
FAR = 1%	83.77	88.07	87.97	88.43	90.43	**90.86**
FAR = 0.1%	55.43	70.97	68.33	72.67	80.60	**81.20**

Bold values indicate the best result

The verification metrics are calculated using LightCNN

best verification performance. It is worth noting that our method can improve the verification performance by a large margin for very low input-resolutions compared with other methods. For example, our method can improve the values of TPR@FAR = 1% on 16×16 and 8×8 input-resolutions to 90.86% and 42.87%, respectively. This fact illustrates the strong ability of our method to recover identity information from very low-resolution faces.

4.6 Ablation Study

In this section, we conduct the ablation study to gain insight into the respective roles of each part of our model in face hallucination. We take the case of hallucinating the LR faces of 16×16 input-resolution in the LFW dataset for example and report the qualitative and quantitative results in Fig. 10 and Table 4. ImgRecNet and ImgAdvNet are two variant nets to show the effectiveness of wavelet-domain generator and discriminator. Our-CNN, Our-CNN+adv and Our-CNN+id are the models trained with different combinations of the loss functions by setting the hyper-parameters $\alpha_2 = \alpha_3 = 0$, $\alpha_3 = 0$ and $\alpha_2 = 0$, respectively. Details are discussed in the following.

Wavelet-Domain Generator To demonstrate the effectiveness of our method comes from the proposed wavelet-domain method rather than the deep network architecture, we replace the wavelet prediction net in Fig. 5 with a sequence of de-convolution blocks, of which the architecture is the same with the generator of DCGAN (Radford et al. 2016) except the input is the output feature maps of the embedding net. The new network is denoted as ImgRecNet and trained using a traditional MSE loss in image pixel domain. From Fig. 10 and Table 4, we can see that wavelet-domain generator Our-CNN outperforms image-pixel-domain generator ImgRecNet both qualitatively and quantitatively, which

Table 5 The influence of different classification net. The baseline model is WaveletSRCNN trained without the identity-preserving loss

Test	Metric	Train						
		Original	Bicubic	Baseline	VGG	VGGFace	LightCNN	LightCNN-29v2
VGGFace	AUC	99.33	86.61	97.49	97.17	97.94	**98.07**	97.86
	FAR = 1%	89.90	24.77	72.57	68.47	74.13	**77.50**	73.97
	FAR = 0.1%	79.63	6.93	45.46	47.03	46.77	**55.87**	47.20
LightCNN	AUC	99.31	90.68	98.07	97.48	98.51	98.48	**98.54**
	FAR = 1%	97.77	45.50	88.43	85.70	90.73	90.86	**91.43**
	FAR = 0.1%	96.23	21.17	72.67	70.63	80.93	**81.20**	76.27
LightCNN-29v2	AUC	99.47	95.65	98.69	98.50	98.97	**99.07**	99.04
	FAR = 1%	99.53	64.30	91.57	87.93	94.13	94.20	**94.60**
	FAR = 0.1%	99.30	43.73	81.87	71.10	85.76	85.87	**87.43**

Bold values indicate the best result

Table 6 The running time (ms) on GPU and CPU

Hardware	Settings	Bicubic	WTIP	SRCNN	SRGAN	SRDense	CBN	URDGN	TDN	TDAE	Ours
GPU	32×32, 4×	–	–	3.805	15.12	25.24	1890	–	–	–	9.306
	16×16, 8×	–	–	3.805	6.259	16.20	2067	13.45	19.11	39.13	9.001
	8×8, 16×	–	–	3.805	3.916	19.83	2128	–	–	–	33.93
CPU	32×32, 4×	0.8903	22.50	223.0	627.1	1051	–	–	–	–	1263
	16×16, 8×	0.8823	22.64	223.0	427.9	315.2	-	1268	1306	2765	700.8
	8×8, 16×	0.8765	19.85	223.0	394.4	149.7	3840[a]	–	–	–	975.9

[a]The run time of CBN on CPU is copied from the cited paper for no publicly available cpu version of CBN code

illustrates that wavelet domain is more suitable than image pixel domain to deal with face hallucination.

Wavelet-Domain Discriminator To show the effectiveness of the wavelet-domain discriminator WaveletDNet, we also train the generator Our-CNN adversarially against an image-pixel-domain discriminator, i.e., the discriminator of SRGAN (Ledig et al. 2017) here. The discriminator of the new GAN model (named as ImgAdvNet) takes the generated and real HR images as inputs and tries to distinguish them. To be fair, we train ImgAdvNet the same epochs with Our-CNN+adv, i.e., 30 epochs. As shown in Fig. 10 and Table 4, Our-CNN+adv with wavelet-domain discriminator achieves better visual perceptual and verification performances than ImgAdvNet with image-pixel-domain discriminator, though the latter has a little higher PSNR and SSIM values. This demonstrates that adversarial learning in wavelet domain can capture more local texture details and generate more perceptually plausible images, which is also helpful to face recognition.

Wavelet Adversarial Loss From Fig. 10, it can be seen that Our-CNN+adv has more texture details than Our-CNN, such as hair, beard and tooth textures. From Table 4, it can be seen that Our-CNN+adv achieves lower PSNR and SSIM values while better verification performance than Our-CNN. This again proves that adversarial learning in wavelet domain can generate perceptually plausible results with abundant texture details and recover more identity information. The comparison between Our-CNN+id and the proposed method can also support this argument.

Identity Preserving Loss From Fig. 10, we can see that the results of Our-CNN+id look more similar with the ground-truth compared with the other methods without the identity preserving loss. The verification measures in Table 4 show that the addition of the identity preserving loss can improve face verification performance by a large margin, while it also causes a little decrease in PSNR and SSIM values like the adversarial loss. Both the visual perceptual and quantitative results demonstrate that the identity preserving loss is essential to improve face recognition performance via face hallucination.

To sum up, we can conclude that the wavelet-domain generator with the wavelet reconstruction loss is superior to the image-pixel-domain generator with the image reconstruction loss (i.e., MSE loss in image pixel domain); the wavelet-domain discriminator with the wavelet adversarial loss can generate photo-realistic texture details and is helpful to face verification; the identity preserving loss can bring a great improvement to face verification via face hallucination.

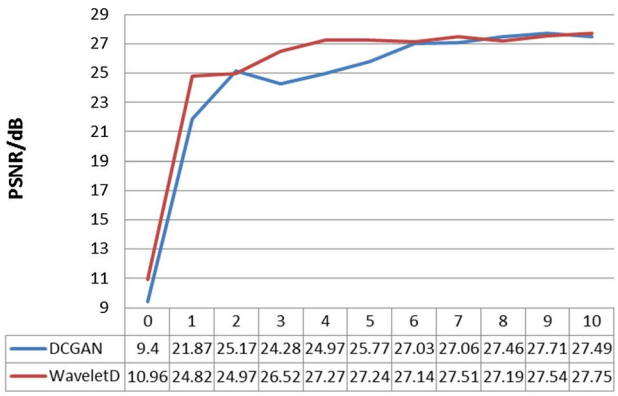
Fig. 11 The visual results of different discriminators along with the training epochs. **a** and **b** are the input and ground-truth images, respectively. **c**–**m** are the results after the training epochs 0–10, respectively. The top three and bottom three rows are the results of WaveletCNN training with the original discriminator of DCGAN and the proposed wavelet-domain discriminator, respectively

4.7 EvalNet Analysis

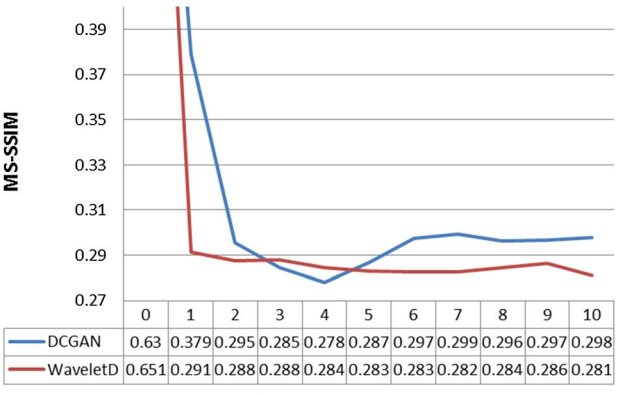
Extensive experiments are conducted on CelebA faces of pixel-size 16×16 with magnification factor $8\times$ to study the influence of different evaluation networks on face verification performance. We train and evaluate the proposed model with four classification networks, i.e., VGG (Simonyan and Zisserman 2015), VGGFace(Parkhi et al. 2015), LightCNN (Wu et al. 2018) and LightCNN-29v2 (the newest and strongest version of LightCNN).

As shown in Table 5, testing with a better face recognition network always achieves better performance, while different training settings have significant influences on the performance. Compared to Bicubic interpolation, the baseline model trained with no classification network has already improved the face verification performance greatly. However, the performance degrades if using the pre-trained VGG network, which implies that non-face classification network may not be helpful for face hallucination on face verification performance. When training with different face recognition nets, such as VGGFace, LightCNN and LightCNN-29v2, the face verification performances improve with a large margin but they are very close to each other, especially when testing with a better classification net like LightCNN-29v2.

In summary, the face verification performance of the proposed method may improve with a face recognition network while degrade with a non-face classification network, but it is not very sensitive to the specific selection of the face recognition network.



(a) PSNR



(b) MS-SSIM

Fig. 12 Comparison on PSNR (higher is better) and MS-SSIM (lower is better) along with the training epochs

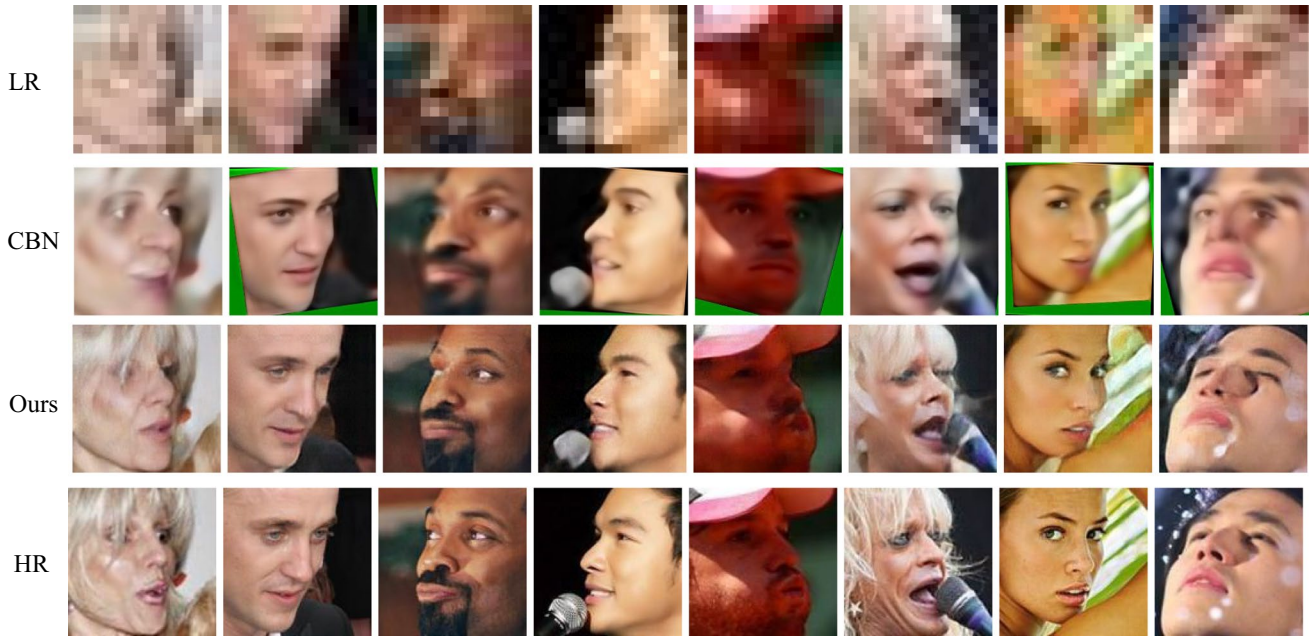


Fig. 13 Robustness toward large pose variations on 16×16 input resolution with $8\times$ upscaling factor. All the images are selected from the CelebA testset

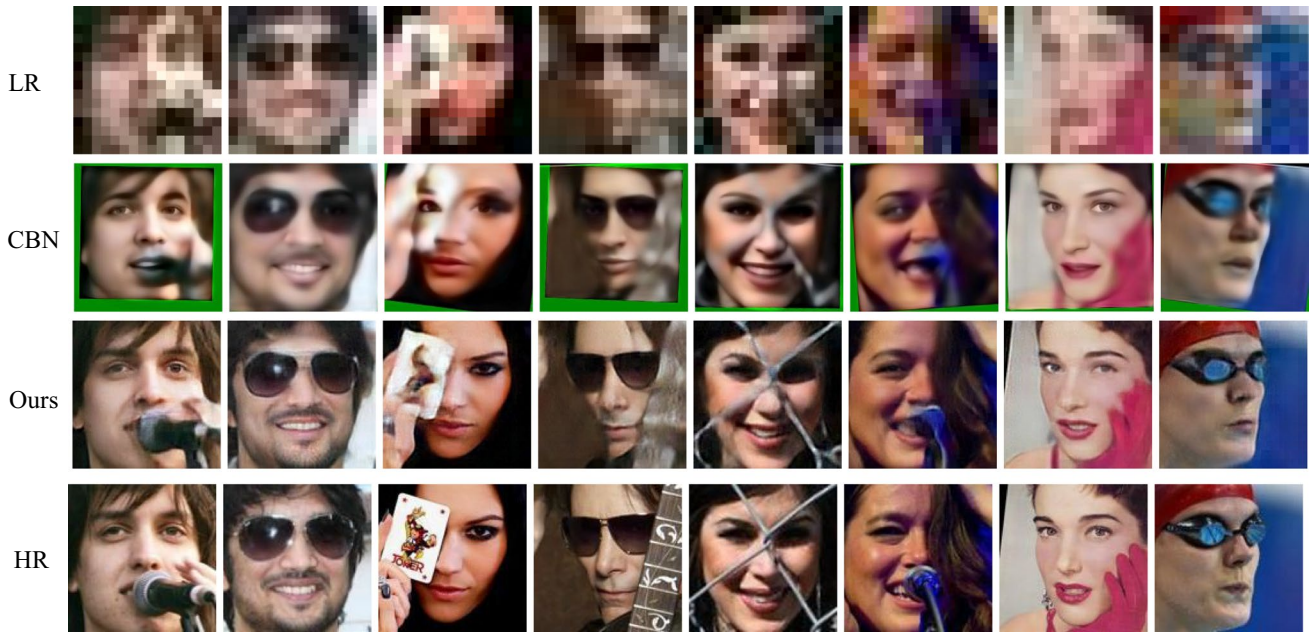


Fig. 14 Robustness toward occlusion variations on 16×16 input resolution with $8\times$ upscaling factor. All the images are selected from the CelebA testset

4.8 Time Complexity

We compare the time complexity against the state-of-the-art models on a single GPU (TITAN Xp GP102) and a single CPU (i7-4790), respectively. As demonstrated in Table 6, the proposed method achieves appealing time performance among the state-of-the-art deep models. It runs the second and third fast on GPU for the $(32 \times 32, 4\times)$ and $(16 \times 16, 8\times)$ settings, respectively; its time cost increases a lot for the $(8 \times 8, 16\times)$ setting but is still less than CBN's. When running on CPU, our method is faster than CBN, URDGN, TDN and TDAE. The superior performance on GPU comes from that the presented network consists of multiple independent subnets, i.e., the subnets in the wavelet prediction net.

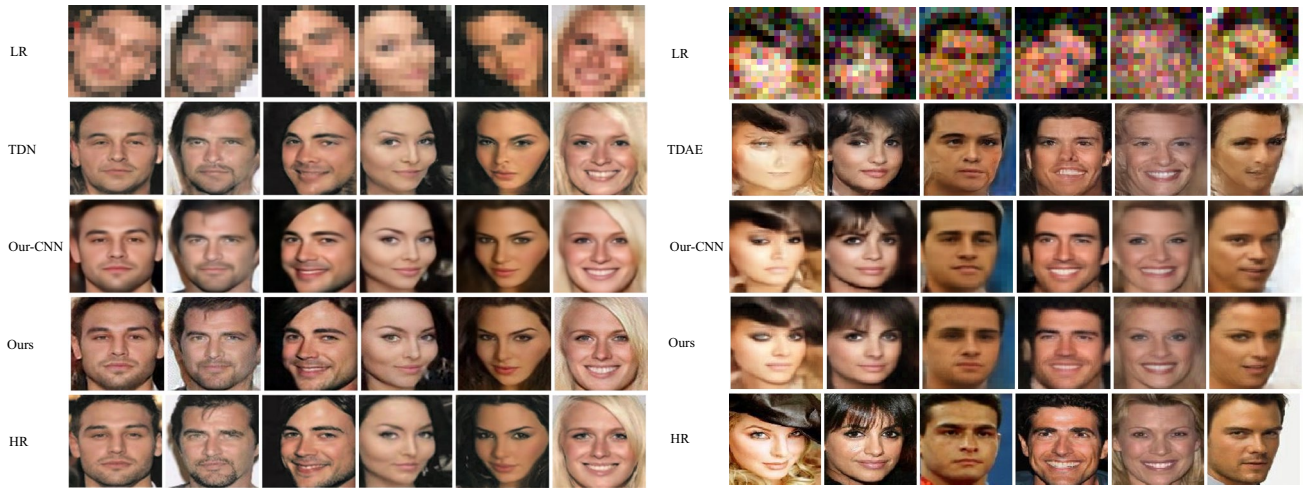


Fig. 15 Comparison on 16 × 16 unaligned faces with/without 10% Gaussian noise. The top two and the bottom rows in left panel are copied from the cited paper

Table 7 Quantitative results on unaligned noisy data

Input image	Method	CelebA		LFW				
		PSNR	SSIM	PSNR	SSIM	AUC	FAR = 1%	FAR = 0.1%
16 × 16, unaligned	TDN	22.66[a]	0.66[a]	–	–	–	–	–
	Our-CNN	**23.31**	**0.8349**	**22.70**	**0.8373**	97.53	81.13	66.50
	Ours	22.39	0.7913	22.03	0.7973	**97.84**	**86.40**	**68.03**
16 × 16, unaligned, noisy	TDAE	20.20 (20.47[a])	**0.8525** (0.56[a])	19.31	0.7153	76.48	15.27	**5.667**
	Our-CNN	20.40	0.7491	19.30	0.7258	75.48	11.73	2.900
	Ours	**20.80**	0.7574	**19.67**	**0.7382**	**79.06**	**17.50**	5.33
8 × 8, unaligned	Our-CNN	**21.29**	**0.7727**	**20.99**	**0.7777**	84.97	27.00	9.867
	Ours	20.82	0.7392	20.56	0.7462	**87.75**	**32.60**	**19.60**
8 × 8, unaligned, noisy	Our-CNN	**18.35**	0.6806	17.59	0.6637	60.45	3.267	**0.6333**
	Ours	18.71	**0.6913**	**17.95**	**0.6767**	**62.55**	**4.100**	**0.6333**

Bold values indicate the best result

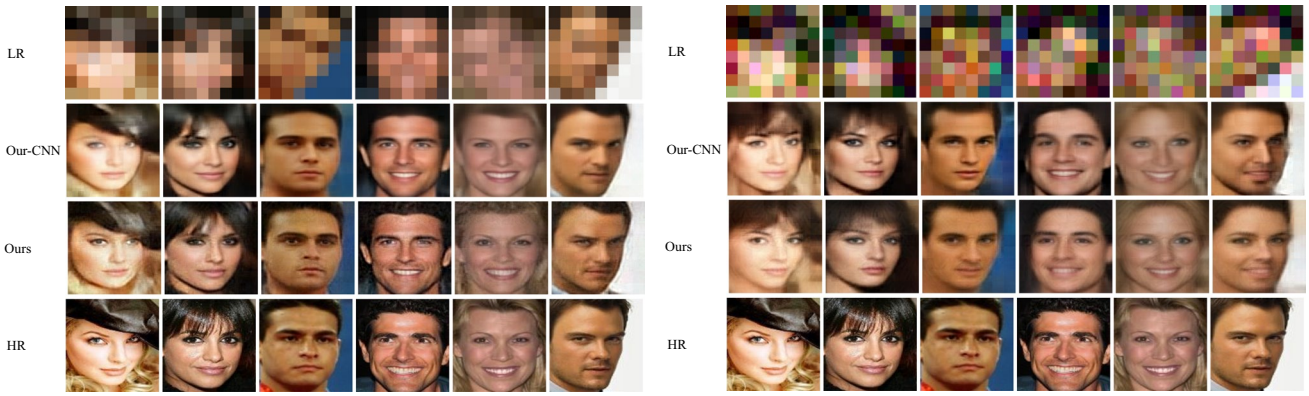
[a]Means the results are copied from the cited papers. The different PSNRs and SSIMs for TDAE come from the different test-sets and the different computation details of SSIM

This speeds up the execution time very much when running parallelly on GPU.

4.9 Training Stability

The training instability still remains a challenge to generative adversarial networks (GANs), especially for high-resolution images. Different from most of the current GANs, we propose a wavelet domain discriminator (WaveletDNet) to synthesize realistic wavelets. Since each wavelet has the same small size with the low-resolution inputs, a very shallow network architecture is able to ensure enough receptive field for the adversarial learning. This reduces the difficulty in training GANs for high-resolution images.

We conduct experiments to compare the training stability, where two different discriminators are used to train WaveletSRCNN, i.e., an image domain discriminator used in DCGAN and the proposed WaveletDNet. As shown in Fig. 11, the proposed WaveletDNet converges faster than the image discriminator and the image results are more appealing. Besides, we adopt two metrics, i.e., PSNR and multi-scale structural similarity (MS-SSIM) (Odena et al. 2017), to quantitatively evaluate the training stability. As shown in Fig. 12, WaveletDNet outperforms the image discriminator in training stability for both the metrics.



(a) Unaligned faces **(b)** Unaligned noisy faces

Fig. 16 Results on 8×8 unaligned faces with/without 10% Gaussian noise

4.10 Robustness Towards Poses and Occlusions

We evaluate the robustness of our method toward large pose and occlusion variations. Some visual results of our method compared with the state-of-the-art CBN on the LR faces with large poses and occlusions are reported in Figs. 13 and 14.

Large Poses As shown in Fig. 13, CBN fails to reconstruct plausible HR faces of large poses, where the edges are blurry and the hidden facial parts are synthesized like ghost. This is because CBN incorporates corresponding field estimation with face hallucination, which is hard to be accurate with large poses. Meanwhile, our method can still infer plausible high-quality images with abundant textures. This demonstrates that dealing with face hallucination in wavelet-domain helps to infer high-frequency details while maintaining plausible global facial topology.

Occlusions We take some faces with natural occlusions for example. As shown in Fig. 14, as the accurate location of the landmarks is hard for the occluded LR faces, CBN tends to over-synthesize occluded facial parts and generate blurry edges around the occlusion boarders. Different from CBN, our method super-resolves the occluded and the rest facial parts dependently and is able to produce high-quality images, not only for the facial parts but also for the occlusions.

Besides, since the LFW dataset contains large quantities of face images captured in the wild, which includes large variations of poses, occlusions and other noises, the appealing evaluation results in the former sections provide additional beneficial evidence that the proposed method has promising robustness toward unconstrained face images with large poses, occlusions and so on.

4.11 Hallucinating Unaligned and Noisy Faces

We conduct experiments to explore the performance of the proposed method for unaligned faces with/without noise. Following the protocols of TDN (Yu and Porikli 2017a) and TDAE (Yu and Porikli 2017b), we train the proposed network to predict the aligned high-resolution faces directly from the unaligned (noisy) low-resolution inputs.

As demonstrated in Fig. 15 and Table 7, the performance of the proposed method on the unaligned faces of 16×16 pixels is comparable to or better than that of TDN/TDAE both quantitatively and qualitatively. We also evaluate the proposed method on the unaligned faces of 8×8 pixels. It can be seen from Fig. 16 that our method is able to predict high-resolution faces from 8×8 unaligned faces while preserving most of the facial information. However, it fails to preserve the facial information for 8×8 unaligned faces with 10% Gaussian noise.

The face verification results in Table 7 demonstrate that it is still difficult to recover the identity information from the unaligned tiny faces polluted by a large amount of noise. Super-resolving unaligned noisy tiny faces remains an open and challenging task.

5 Conclusion

We propose a novel wavelet-domain generative adversarial approach for multi-scale face hallucination, which transforms single image super-resolution to wavelet coefficients prediction task in deep learning framework. A flexible wavelet-domain generative adversarial network (WaveletSRGAN) is presented, which consists of three subnetworks: wavelet-domain super-resolution network , wavelet-domain discriminator network and facial evaluation network. Three types of losses, i.e., wavelet reconstruction loss, wavelet adversarial loss and identity preserving loss, are designed to generate abundant photo-realistic texture details while maintaining the global facial

topology information. Due to its extensible fully convolutional architecture trained with simply-aligned faces, our network is applicable to different input resolutions with various upscaling factors. Experimental results show that the proposed method demonstrates promising robustness toward very low-resolution faces with large pose and occlusion variances. It achieves more appealing results both qualitatively and quantitatively than the state-of-the-arts, and can significantly improve identity verification performance for low-resolution face images captured in the wild.

Acknowledgements This work is partially funded by the State Key Development Program (Grant No. 2016YFB1001001), National Natural Science Foundation of China (Grant No. 61622310, 61427811), and Beijing Natural Science Foundation (Grants No. JQ18017).

References

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

Bruna, J., Sprechmann, P., & LeCun, Y. (2016). Super-resolution with deep convolutional sufficient statistics. In *International conference on learning representations*.

Bulat, A., & Tzimiropoulos, G. (2018). Super-fan: Integrated facial landmark localization and super-resolution of real-world low resolution faces in arbitrary poses with GANS. In *IEEE conference on computer vision and pattern recognition* (pp. 109–117).

Bulat, A., Yang, J., & Tzimiropoulos, G. (2018). To learn image super-resolution, use a GAN to learn how to do image degradation first. In *European conference on computer vision* (pp. 185–200).

Chang, H., Yeung, D. Y., & Xiong, Y. (2004). Super-resolution through neighbor embedding. In *IEEE Conference on Computer Vision and Pattern Recognition* (Vol. 1, pp. 275–282).

Chen, Y., Tai, Y., Liu, X., Shen, C., & Yang, J. (2018). FSRNet: End-to-end learning face super-resolution with facial priors. In *IEEE conference on computer vision and pattern recognition* (pp. 2492–2501).

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

Dahl, R., Norouzi, M., & Shlens, J. (2017). Pixel recursive super resolution. In *IEEE international conference on computer vision* (pp. 5439–5448).

Dong, C., Loy, C. C., He, K., & Tang, X. (2016). Image super-resolution using deep convolutional networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *38*(2), 295–307.

Farrugia, R. A., & Guillemot, C. (2017). Face hallucination using linear models of coupled sparse support. *IEEE Transactions on Image Processing*, *26*(9), 4562–4577.

Gao, X., & Xiong, H. (2016). A hybrid wavelet convolution network with sparse-coding for image super-resolution. In *IEEE international conference on image Processing* (pp. 1439–1443).

Gatys, L. A., Ecker, A. S., & Bethge, M. (2016) Image style transfer using convolutional neural networks. In *IEEE conference on computer vision and pattern recognition* (pp. 2414–2423).

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., et al. (2014). Generative adversarial nets. *Advances in Neural Information Processing Systems*, *27*, 2672–2680.

Hayat, M., Khan, S. H., & Bennamoun, M. (2017). Empowering simple binary classifiers for image set based face recognition. *International Journal of Computer Vision*, *123*(3), 479–498.

Huang, G. B., Ramesh, M., Berg, T., & Learned-Miller, E. (2007) *Labeled faces in the wild: A database for studying face recognition in unconstrained environments*. Technical Report 07-49, University of Massachusetts, Amherst.

Huang, H., He, R., Sun, Z., & Tan, T. (2017). Wavelet-SRNet: A wavelet-based CNN for multi-scale face super resolution. In *IEEE international conference on computer vision* (pp. 1689–1697).

Huang, H., Li, Z., He, R., Sun, Z., & Tan, T. (2018). Introvae: Introspective variational autoencoders for photographic image synthesis. In *Neural information processing systems*.

Huang, J. B., Singh, A., & Ahuja, N. (2015). Single image super-resolution from transformed self-exemplars. In *IEEE conference on computer vision and pattern recognition* (pp. 5197–5206).

Ioffe, S., & Szegedy, C. (2015). Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International conference on machine learning* (pp. 448–456).

Ji, H., & Fermüller, C. (2009). Robust wavelet-based super-resolution reconstruction: Theory and algorithm. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *31*(4), 649–660.

Jiang, J., Hu, R., Wang, Z., & Han, Z. (2014). Noise robust face hallucination via locality-constrained representation. *IEEE Transactions on Multimedia*, *16*(5), 1268–1281.

Johnson, J., Alahi, A., & Fei-Fei, L. (2016). Perceptual losses for real-time style transfer and super-resolution. In *European conference on computer vision* (pp. 694–711).

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

Karras, T., Aila, T., Laine, S., & Lehtinen, J. (2018). Progressive growing of GANs for improved quality, stability, and variation. In *International conference on learning representations*.

Kim, J., Kwon Lee, J., & Mu Lee, K. (2016a). Accurate image super-resolution using very deep convolutional networks. In *IEEE conference on computer vision and pattern recognition* (pp. 1646–1654).

Kim, J., Kwon Lee, J., & Mu Lee, K. (2016b). Deeply-recursive convolutional network for image super-resolution. In *IEEE conference on computer vision and pattern recognition* (pp. 1637–1645).

Kingma, D., & Ba, J. (2014). Adam: A method for stochastic optimization. In *International conference on learning representations*.

Lai, W. S., Huang, J. B., Ahuja, N., & Yang, M. H. (2017). Deep Laplacian pyramid networks for fast and accurate super-resolution. In *IEEE conference on computer vision and pattern recognition* (pp. 624–632).

Ledig, C., Theis, L., Huszar, F., Caballero, J., Cunningham, A., Acosta, A., Aitken, A., Tejani, A., Totz, J., Wang, Z., & Shi, W. (2017). Photo-realistic single image super-resolution using a generative adversarial network. In *IEEE conference on computer vision and pattern recognition* (pp. 4681–4690).

Li, B., Chang, H., Shan, S., & Chen, X. (2009). Aligning coupled manifolds for face hallucination. *IEEE Signal Processing Letters*, *16*(11), 957–960.

Lin, Z., He, J., Tang, X., & Tang, C. K. (2008). Limits of learning-based superresolution algorithms. *International Journal of Computer Vision*, *80*(3), 406–420.

Liu, C., Shum, H. Y., & Freeman, W. T. (2007). Face hallucination: Theory and practice. *International Journal of Computer Vision*, *75*(1), 115–134.

Liu, Z., Luo, P., Wang, X., & Tang, X. (2015). Deep learning face attributes in the wild. In *IEEE international conference on computer vision* (pp. 3730–3738).

Ma, X., Zhang, J., & Qi, C. (2010). Hallucinating face by position-patch. *Pattern Recognition*, *43*(6), 2224–2236.
Mallat, S. (1996). Wavelets for a vision. *Proceedings of the IEEE*, *84*(4), 604–614.
Mallat, S. (2016). Understanding deep convolutional networks. *Philos Trans R Soc A*, *374*(2065), 20150203.
Mallat, S. G. (1989). A theory for multiresolution signal decomposition: The wavelet representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *11*(7), 674–693.
Mao, X., Li, Q., Xie, H., Lau, R. Y., Wang, Z., & Smolley, S. P. (2017). Least squares generative adversarial networks. In *IEEE international conference on computer vision* (pp. 2813–2821).
Naik, S., & Patel, N. (2013). Single image super resolution in spatial and wavelet domain. *The International Journal of Multimedia & Its Applications*, *5*(4), 23.
Nguyen, N., & Milanfar, P. (2000). A wavelet-based interpolation-restoration method for superresolution (wavelet superresolution). *Circuits, Systems, and Signal Processing*, *19*(4), 321–338.
Odena, A., Olah, C., & Shlens, J. (2017). Conditional image synthesis with auxiliary classifier GANs. In *International conference on machine learning* (pp. 2642–2651).
van den Oord, A., Kalchbrenner, N., Espeholt, L., kavukcuoglu, k, Vinyals, O., & Graves, A. (2016). Conditional image generation with pixelcnn decoders. *Advances in Neural Information Processing Systems*, *29*, 4790–4798.
Park, J. S., & Lee, S. W. (2008). An example-based face hallucination method for single-frame, low-resolution facial images. *IEEE Transactions on Image Processing*, *17*(10), 1806–1816.
Parkhi, O. M., Vedaldi, A., & Zisserman, A. (2015). Deep face recognition. In *British machine vision conference*.
Radford, A., Metz, L., & Chintala, S. (2016). Unsupervised representation learning with deep convolutional generative adversarial networks. In *International conference on learning representations*.
Sajjadi, M. S. M., Scholkopf, B., & Hirsch, M. (2017). Enhancenet: Single image super-resolution through automated texture synthesis. In *IEEE international conference on computer vision* (pp. 4491–4500).
Shamir, L. (2008). Evaluation of face datasets as tools for assessing the? Performance of face recognition methods. *International Journal of Computer Vision*, *79*(3), 225.
Shi, W., Caballero, J., Huszár, F., Totz, J., Aitken, AP., Bishop, R., Rueckert, D., & Wang, Z. (2016). Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network. In *IEEE conference on computer vision and pattern recognition* (pp. 1874–1883).
Simonyan, K., & Zisserman, A. (2015). Very deep convolutional networks for large-scale image recognition. In *International conference on learning representations*.
Singh, A., Porikli, F., & Ahuja, N. (2014). Super-resolving noisy images. In *IEEE conference on computer vision and pattern recognition* (pp. 2846–2853).
Sohn, K,. Liu, S., Zhong, G., Yu, X., Yang, M. H., & Chandraker, M. (2017). Unsupervised domain adaptation for face recognition in unlabeled videos. In *IEEE international conference on computer vision* (pp. 3210–3218).
Sønderby, C. K., Caballero, J., Theis, L., Shi, W., & Huszár, F. (2017). Amortised map inference for image super-resolution. In *International conference on learning representations*.
Sun, J., Xu, Z., & Shum, H. Y. (2008). Image super-resolution using gradient profile prior. In *IEEE conference on computer vision and pattern recognition* (pp. 1–8).
Tai, Y., Yang, J., & Liu, X. (2017). Image super-resolution via deep recursive residual network. In *IEEE conference on computer vision and pattern recognition* (pp. 3147–3155)
Tong, T., Li, G., Liu, X., & Gao, Q. (2017). Image super-resolution using dense skip connections. In *IEEE international conference on computer vision* (pp. 4799–4807).
Wang, N., Tao, D., Gao, X., Li, X., & Li, J. (2014). A comprehensive survey to face hallucination. *International Journal of Computer Vision*, *106*(1), 9–30.
Wang, X., & Tang, X. (2005). Hallucinating face by eigentransformation. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, *35*(3), 425–434.
Wu, X., Song, L., He, R., & Tan, T. (2018). Coupled deep learning for heterogeneous face recognition. In *AAAI conference on artificial intelligence*.
Xu, X., Sun, D., Pan, J., Zhang, Y., Pfister, H., & Yang, M. H. (2017). Learning to super-resolve blurry face and text images. In *IEEE international conference on computer vision* (pp. 251–260).
Yang, C. Y., & Yang, M. H. (2013). Fast direct super-resolution by simple functions. In *IEEE international conference on computer vision* (pp. 561–568)
Yang, C. Y., Liu, S., & Yang, M. H. (2013) Structured face hallucination. In *IEEE conference on computer vision and pattern recognition* (pp. 1099–1106).
Yang, C. Y., Liu, S., & Yang, M. H. (2017). Hallucinating compressed face images. *International Journal of Computer Vision*. https://doi.org/10.1007/s11263-017-1044-4.
Yang, J., Tang, H., Ma, Y., & Huang, T. (2008). Face hallucination via sparse coding. In *IEEE international conference on image processing* (pp. 1264–1267).
Yang, J., Wright, J., Huang, T. S., & Ma, Y. (2010). Image super-resolution via sparse representation. *IEEE Transactions on Image Processing*, *19*(11), 2861–2873.
Yu, X., & Porikli, F. (2016). Ultra-resolving face images by discriminative generative networks. In *European conference on computer vision* (pp. 318–333).
Yu, X., & Porikli, F. (2017a). Face hallucination with tiny unaligned images by transformative discriminative neural networks. In *AAAI conference on artificial intelligence* (pp. 4327–4333).
Yu, X., & Porikli, F. (2017b). Hallucinating very low-resolution unaligned and noisy face images by transformative discriminative autoencoders. In *IEEE conference on computer vision and pattern recognition* (pp. 3760–3768).
Yu, X., Fernando, B., Ghanem, B., Porikli, F., & Hartley, R. (2018a). Face super-resolution guided by facial component heatmaps. In *European conference on computer vision* (pp. 217–233).
Yu, X., Fernando, B., Hartley, R., & Porikli, F. (2018b). Super-resolving very low-resolution face images with supplementary attributes. In *IEEE conference on computer vision and pattern recognition* (pp. 908–917).
Zhang, H., Xu, T., Li, H., Zhang, S., Wang, X., Huang, X., & Metaxas, DN. (2017). Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks. In *IEEE international conference on computer vision* (pp. 5907–5915).
Zhao, S., Han, H., & Peng, S. (2003). Wavelet-domain HMT-based image super-resolution. *IEEE International Conference on Image Processing*, *2*, 953–956.
Zhu, S., Liu, S., Loy, C. C., & Tang, X. (2016). Deep cascaded bi-network for face hallucination. In *European conference on computer vision* (pp. 614–630).