This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMM.2019.2898752, IEEE Transactions on Multimedia

Face Hallucination via Coarse-to-Fine Recursive Kernel Regression Structure

Jingang Shi and Guoying Zhao *Senior Member, IEEE*

***Abstract*—In recent years, patch-based face hallucination algorithms have attracted considerable interest due to their effectiveness. These approaches produce a high-resolution (HR) face image according to the corresponding low-resolution (LR) input by learning a reconstruction model from the given training image set. The critical problem in these algorithms is establishing the underlying relationship between LR and HR patch pairs. Most previous methods aim to denote each input LR patch by the linear combination of the training set in the LR space while utilizing the combination weights to reconstruct the target HR patch. However, this assumes that the same combination weights should be shared between various resolution spaces, which is truly difficult to satisfy because of the one-to-many mapping relation between LR and HR patches. In this paper, we directly train a series of adaptive kernel regression mappings for predicting the lost high-frequency information from the LR patch, which avoids dealing with the above difficult problem. During the training process, we first establish a local optimization function on each LR/HR training pair according to the geometric structure of neighboring patches. The objective of local optimization can be presented in two aspects: (1) ensure the reconstruction consistency between each LR patch and the corresponding HR patch; (2) preserve the intrinsic geometry between each HR training patch and its original neighbors after the reconstruction process. The local optimizations are finally incorporated as the global optimization for calculating the optimal kernel regression function. To better approximate the target HR patch, we further propose a recursive structure to compensate for the residual reconstruction error of high-frequency details by a series of regression mappings. The proposed method is rather fast yet very effective in producing HR face images. Experimental results show that the proposed approach achieves superior performance with reasonable computational time compared with the state-of-the-art methods.**

***Index Terms*—Face hallucination, low-resolution, patch-based, kernel regression, super-resolution.**

I. Introduction

FACE image is an important human feature which has been widely applied in the field of computer vision and pattern recognition. Consequently, a growing number of applications have been developed and investigated according to face image analysis, which includes detection, alignment, tracking and recognition. These techniques play an important role in human identification, security control, surveillance monitoring and digital entertainment. However, the performance of most face analysis tasks degrades significantly when the target face images are captured under uncontrolled conditions. Specifically, the variation in image resolution is a critical factor in practical scenarios in which high-quality digital cameras are not deployed because of the limitation of cost and storage space.

Low-resolution images truly cause many restrictions in the real-world applications with high definition requirements. It is desirable to reconstruct high-resolution images from low-resolution images by super-resolution (SR) algorithms [1]. According to recent SR research, these methods are classified into two types: interpolation-based approaches [2]–[5] and learning-based approaches [6]–[43]. The interpolation-based approaches estimate statistical prior knowledge from natural images to produce HR images, but there are inherent limitations when dealing with the increase in the magnification factor. In contrast, the learning-based approaches reconstruct the final HR results with a set of LR/HR training samples. Due to the extra information from the training set, the learning-based approaches have the ability to achieve better visual quality and deal with higher magnification factors. In this paper, we mainly discuss the learning-based SR algorithms.

The pioneering work on face hallucination was proposed by Baker and Kanade [11]. Depending on the advantages of learning-based methods, the output HR face image can be inferred from the corresponding LR input by a Bayesian formulation. Due to the potential applications in face analysis, face hallucination has become an active area of research. Many face hallucination methods have been introduced to produce HR images by various reconstruction models. Early studies aimed to utilize a two-step procedure to hallucinate LR face images. It first produced a global face image with plausible contour but lacked vivid textures. The local details were further compensated in patch-wise. Liu et al. [15] learned a global parametric Gaussian model to describe the relationship between LR face images and corresponding HR images, while the details were enhanced by local patches using a nonparametric Markov random field (MRF) model. Zhuang et al. [16] estimated the relationship between LR and HR images by locality preserving projection and refined it by radial basis function regression. Details of the hallucinated face image were further enhanced by neighbor embedding [12]. Park and Lee [14] reconstructed global face images by an extended morphable model and further compensated them with an error back-propagation method. Jia and Gong [17] utilized a hierarchical tensor representation to achieve hallucinated face images across multiple modalities. A local patch-based

J. Shi is with the Center for Machine Vision and Signal Analysis, University of Oulu, FI-90014 Oulu, Finland. G. Zhao is with the School of Information and Technology, Northwest University, Xi'an 710069, China and with the Center for Machine Vision and Signal Analysis, University of Oulu, FI-90014 Oulu, Finland. E-mail: (jingang.shi@oulu.fi; guoying.zhao@oulu.fi).(Corresponding author: Guoying Zhao)

This work was funded by the National Natural Science Foundation of China (No. 61772419), Academy of Finland, Bussiness Finland, Tekniikan Edistämissäätiö and Infotech Oulu.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMM.2019.2898752, IEEE Transactions on Multimedia

multiresolution tensor was also designed to generate the local details. Huang and He [18] applied canonical correlation analysis (CCA) to maximize the correlation between global LR and HR images. In the residual compensation phase, the same model was also utilized to find a proper subspace for patches in different resolutions. An and Bhanu [20] further extended the method in [18] by 2D-CCA, which achieved better performance for preserving the global geometric structure and local texture. The disadvantage of these algorithms is that a large number of training samples is required to describe the features of global face images in the corresponding subspace. The performance may not be appropriate when the input LR images share fewer common features with the training set due to the disturbance of pose, illumination or noise, especially for applications in real-world conditions. Usually, the artifacts caused by global reconstruction can not be completely suppressed in the compensation procedure of local details.

Considering the position information of image patches as a constraint, Ma et al. [25] reconstructed HR face images based on position-patches instead of the complicated two-step model and further extended the method to adapt the application across multiple variations [26]. From then on, a number of face hallucination models began to produce HR face images using training patches from the same position. Li et al. [24] trained a couple of projections which transformed the LR and HR patches into a latent feature subspace for estimating the combination relationship. Huang and Wu [21] learned multiple local linear transformations to approximate the regression between LR and HR patches. Jung et al. [27] introduced an ℓ_1-norm minimization constraint for improving the under-determined problem in [25], which produces more stable face SR results. Zhang and Cham [23] predicted the local feature of face images in the discrete cosine transform (DCT) domain. Shi et al. [32] trained position-based dictionaries for sparsely representing local patches to enhance the high-frequency details. Wang et al. [28] incorporated a weighted adaptive sparse regularization method to infer the combination weights of HR patches. Jiang et al. [30] utilized the local similarity constraint for regularizing the HR face reconstruction procedure. Zeng and Huang [33] augmented the existing training set to improve the quality of the final reconstructed results. In [22], Jiang et al. represented the relationship of LR and HR patches by a smooth regression model. The target HR patch was then reconstructed by the corresponding LR patch through weighted linear mapping. Liu et al. [29] devised a weight vector to subtly tune the contribution of examples, which makes the algorithm more robust to the influence of noise.

Due to the development of neural networks, many researchers have taken advantage of deep learning to conduct super-resolution and face hallucination tasks. Dong et al. [34] first solved the SR problem by using a neural network with three convolutional layers to learn the mapping function between LR and HR images. Kim et al. [35] introduced a very deep convolutional network for reconstructing HR images, which significantly improves SR performance. Yu et al. [36] investigated the probability of utilizing a generative adversarial network to produce HR face images with better visual effects. In [37], a transformative discriminative autoencoder was proposed to deal with unaligned and noisy face images. Zhu et al. [38] proposed a gated deep bi-network for face hallucination, which localizes LR facial components and exploits a face spatial prior to recover reasonable details. Song et al. [39] generated an initial HR face image by a convolutional neural network and further followed with an enhancement procedure to refine the results. Cao et al. [40] employed an attention-aware mechanism in the face hallucination task, which selects preferable facial components by reinforcement learning and further reconstructs the local details by an enhancement network. Chen et al. [41] considered the utilization of facial landmark heatmaps and parsing maps in the training phase for obtaining better super-resolving results. Huang et al. [42] proposed a wavelet-based neural network to conduct the face hallucination task, which simultaneously takes into account the local texture details and global topology information of human faces. Yu et al. [43] proposed a multi-task network that consists of upsampling branch and facial component heatmaps estimation branch. The upsampling branch is guided by the heatmaps in the reconstruction for preserving face structure.

Generally, most previous position-patch-based methods [24]–[31] aim to estimate reasonable combination weights which represent each input LR patch as a linear combination of LR training patches. The combination relationship is employed to produce the predicted HR patch according to the help of corresponding HR training set. However, these methods require a basis assumption about sharing the same combination weights between HR and LR spaces. As shown in [44], this strong assumption is hardly satisfied in the application. Because of the one-to-many mapping relation between HR and LR patches, the relationship captured from the LR space cannot present the real situations of the HR space. Though many algorithms have been proposed to devise various priors for improving the tough problem, it is also challenging to choose a suitable regularization model for better SR results.

Another category of position-patch-based face hallucination methods [21] [22] considers the reconstruction of HR patches as a regression problem. These methods aim to learn regression mappings from the training set, which can be employed for directly projecting the LR testing patches into HR patches. The advantage is that these algorithms do not require the assumption of combination relationship consistency between LR and HR spaces. However, it is also very challenging to train such a projection function since the SR reconstruction problem is highly ill-conditioned. It is expected to further improve the regression mapping between LR and HR patches to achieve better reconstruction results.

In the literature [21] [22], various local linear mappings have been trained to describe the regression relationship between LR and HR patches. Nonetheless, it is a limitation to represent the reconstruction procedure as linear regression because of the low dimension of LR feature. Specifically, the whole face image is commonly split into very small overlapped patches to better reflect local details (e.g., only 3×3 for LR case), which means that the regression relationship between LR and HR patches is extremely under-determined. Due to the low dimension of LR feature, the regression is mainly influenced by the additional regularization constraint rather

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMM.2019.2898752, IEEE Transactions on Multimedia

than intrinsic LR feature extracted from the LR patch. To avoid learning the regression relationship from such a low-dimensional feature, we first project the LR patch into a reproducing kernel Hilbert space (RKHS) by kernel trick, which improves the under-determined problem and obtains more stable results. Meanwhile, the kernel function also has the advantage of capturing the nonlinear structures of LR patches. The nonlinear regression is more effective for making a plausible approximation for the one-to-many relationship between LR and HR patches.

Previously, the algorithms aimed to directly predict the projection between LR and HR patches, or implicitly present the relationship by the combination weights of the training patches. None of them considers the original geometry between neighboring patches in the HR space. As we know, the target of most regression-based algorithms [21] [22] is to learn various projection functions with the help of training set, which minimizes the total errors between the estimated HR patches and the original HR patches. However, the local relationship of the HR patches is not further considered in these methods. An illustration is presented in Fig. 1. Though the reconstructed HR patch is well produced to approximate the original patch in the sense of least error, it may fail to preserve the local geometric structure with the neighboring patches. Thus, two image patches that are similar in the original HR space may suffer from some differences after the reconstruction. Actually, the degradation procedure from HR to LR varies the local geometry of image patches, which means that we need to recover the local relationship in the reconstruction. It is no doubt that we can reconstruct plausible HR patches without the constraint of local geometry. However, artifacts may occur in the reconstruction results since the local similarities of neighboring patches are no longer preserved. To improve the above disadvantage, we define local optimization for each training patch according to the corresponding geometric structure. The optimization function not only maintains the reconstruction consistency for each LR patch and corresponding HR patch but also preserves the intrinsic distances between each given sample and its neighbors in the HR space. In this case, the proposed method is feasible to train a reasonable regression function for reconstructing faithful HR patches.

To better refine the target HR patch, we further utilize a coarse-to-fine strategy to compensate for the residual reconstruction error. Since most high-frequency details are lost in the LR image patch, it is hard to directly reconstruct the target HR patch from the degraded LR patch using one regression function. Similar strategies have also been adopted in the literature [45]–[49]. Different from these approaches, the proposed method designs a recursive structure by combining a series of specific kernel regression mappings, which aim to simultaneously preserve the intrinsic geometric structure and eliminate the reconstruction error. Each kernel regression function considers the refined result from the previous mapping as input to predict the residual texture information. The intermediate reconstructed results contain more abundant high-frequency details than the initial LR patch. Consequently, it is suitable to gradually reconstruct the final HR image by such a recursive procedure.

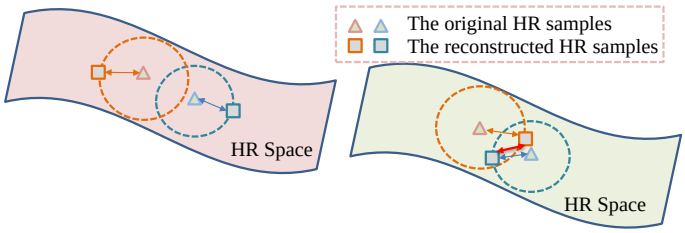


Fig. 1. Without the constraint of local geometry, the relationship between neighboring patches may fail to be preserved. For example, two neighboring HR samples are presented in the above figures, which indicates that they share similar textures in the original HR space. If the reconstruction procedure just considers the consistency of each reconstructed sample and corresponding HR one, regardless of the local distances between neighboring patches in the original HR space, the reconstructed samples may suffer from a relative large difference (such as the example in the left figure). In this case, the two reconstructed patches represent textures with certain diversity, which is not consistent with the real situation in the original HR space. Apparently, the example in the right figure presents a more plausible solution for the reconstruction, which further takes care of the distances between neighboring patches in the optimization.

The main contributions of this paper are presented in the following:

- Motivated by the ability of the kernel function to capture the nonlinear feature of the LR patch, we employ kernel regression mapping for predicting the corresponding HR image patch, which also produces more stable results than the linear projection.
- Different from the previous algorithms that only maintain the reconstruction consistency between each LR patch and the corresponding HR patch, the proposed algorithm also considers the intrinsic geometric structure between each training sample and its neighboring patches. In the reconstruction procedure, it is critical to preserving the distances of the neighboring patches in the HR space to achieve better results.
- To better approximate the target HR patch, we propose a recursive structure by a series of regression mappings to compensate for the residual reconstruction error. The coarse-to-fine strategy further refines the high-frequency details of the output HR image.
- The proposed algorithm is rather fast yet very effective for producing HR face images. Experimental results show that the proposed method achieves superior performance with reasonable computational cost when compared with the state-of-the-art methods.

II. Related Work

As we know, the relationship between the LR observation image and the original HR image can be represented by the following expression:

$$\mathbf{Y}=\mathbf{DBX}+\varepsilon \tag{1}$$

where $\mathbf{Y}$ is the LR face image, $\mathbf{X}$ is the original HR face image, $\mathbf{B}$ is the blurring filter for the HR image, $\mathbf{D}$ denotes the down-sampling procedure and ε is the additive Gaussian white noise. Define $\boldsymbol{I}_L^n$ and $\boldsymbol{I}_H^n$ as the LR and corresponding HR training images respectively, where the index $n = 1, 2 \ldots N$ represents the number of samples in the training set. The task

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMM.2019.2898752, IEEE Transactions on Multimedia

of face hallucination aims to reconstruct a plausible HR face image from the corresponding LR input. Next, we will briefly review some recent works that are relevant to the proposed algorithm.

A. The position-patch based method

Inspired by the fact that patches from the same position of the face image usually share similar textures, Ma et. al. [25] considered the position prior of local patches as a constraint for achieving superior results in the reconstruction. In their method, face images are split into M overlapped patches based on the position information. Let the LR and HR image patches located on position p as $\boldsymbol{l}_p^n$ and $\boldsymbol{h}_p^n$, respectively. According to the position prior, the entire training set is divided into M groups $\{\mathbf{L}_p, \mathbf{H}_p\}_{p=1}^M$, where $\mathbf{L}_p = \left[\boldsymbol{l}_p^1, \ldots, \boldsymbol{l}_p^N\right]$ and $\mathbf{H}_p = \left[\boldsymbol{h}_p^1, \ldots, \boldsymbol{h}_p^N\right]$ represent the corresponding LR and HR training patches on position p. Meanwhile, each LR face image $\mathbf{Y}$ is also split into M overlapped patches to conduct the reconstruction procedure. Each LR image patch located on position p can be represented as $\boldsymbol{y}_p^L$, while it is expected to recover the corresponding HR patch $\boldsymbol{x}_p^H$ with help from the training set. Specifically, the HR patch $\boldsymbol{x}_p^H$ is assumed to be obtained by a linear combination of training patches. According to the assumption of manifold consistency between the LR and HR spaces, the optimal combination coefficients are estimated in the LR space by minimizing the following function:

$$\boldsymbol{\omega}_p = \arg\min_{\boldsymbol{\omega}_p} \left\{ \left\| \boldsymbol{y}_p^L - \mathbf{L}_p \boldsymbol{\omega}_p \right\|_2^2 + \lambda \Omega\left(\boldsymbol{\omega}_p\right) \right\} \tag{2}$$

where $\Omega\left(\boldsymbol{\omega}_p\right)$ represents the regularization term. In [25], the authors utilized an ℓ_2-norm constraint for regularizing the combination weights, while other researchers have further devised various regularization terms to improve the reconstruction results.

B. The linear regression based method

Recently, Jiang et al. [22] proposed a novel regression-based method using the local structure characteristic of face images, which learns a linear mapping function to estimate the HR patch from the corresponding LR patch. For each image patch at position p, the reconstruction procedure can be described as:

$$\boldsymbol{x}_p^H = \mathbf{A} \boldsymbol{y}_p^L \tag{3}$$

where $\mathbf{A}$ is the linear mapping function that corresponds to the position p. The local linear regression can be learned from the training set by:

$$\mathbf{A} = \arg\min_{\mathbf{A}} \left\{ \sum_i w_i \left\| \boldsymbol{h}_p^i - \mathbf{A} \boldsymbol{l}_p^i \right\|_2^2 + \lambda \left\| \mathbf{A} \right\|_F^2 \right\} \tag{4}$$

where $\boldsymbol{l}_p^i$ and $\boldsymbol{h}_p^i$ represent the LR and HR training images, respectively, λ is the regularization parameter, and the corresponding weight w_i is defined by:

$$w_i = 1/(dist(\boldsymbol{y}_p^L, \boldsymbol{l}_p^i))^\alpha \tag{5}$$

Here, $dist(\boldsymbol{y}_p^L, \boldsymbol{l}_p^i)$ represents the squared Euclidean distance between $\boldsymbol{y}_p^L$ and $\boldsymbol{l}_p^i$, while α is the adjustable factor. Thus, the samples that are the most similar to the input LR patch are assigned larger weights in (4), which encourages these similar samples to be privileged in the training phase. Compared with the previous methods [24]–[31], this algorithm avoids the difficult task of preserving manifold consistency in the LR and HR spaces. One disadvantage is that it still requires to repeat the training process for each input sample since the weight w_i is relevant to the input LR patch $\boldsymbol{y}_p^L$, which will be further improved in the proposed method.

III. The Proposed Algorithm

A. Reconstructing the HR patch by kernel regression

According to (3), the relationship between LR patch $\boldsymbol{y}_p^L \in \mathcal{R}^d$ and HR patch $\boldsymbol{x}_p^H \in \mathcal{R}^m$ can be simply represented as a linear regression model by employing the LR training set $\mathbf{L}_p = \left[\boldsymbol{l}_p^1, \ldots, \boldsymbol{l}_p^N\right]$ and HR training set $\mathbf{H}_p = \left[\boldsymbol{h}_p^1, \ldots, \boldsymbol{h}_p^N\right]$. However, the whole face image is usually divided into very small overlapped patches to reflect the local details (e.g., 3×3 pixels), which means that LR patch $\boldsymbol{y}_p^L$ is in a very low dimension. In this case, the linear regression between LR and HR patches is extremely under-determined, which causes difficulty in effectively addressing the complex input-output relationship. To overcome the shortcoming of the linear regression model, we first project the LR features into RKHS by a kernel function [50]. The dimension of the feature in the kernel space is much higher than the original dimension of the LR feature. Thus, in the case of using kernel regression, the project matrix contains more parameters to fit the complex relationship between LR and HR patches. Due to the additional parameters in the kernel regression, this can result in more accurate approximation to the input-output relationship, which makes the face hallucination procedure more stable. Meanwhile, the kernel regression also has the advantage of exploiting the nonlinear characteristics to better conduct the face reconstruction procedure.

Define $\phi : \mathcal{R}^d \rightarrow \mathcal{R}^{\mathcal{F}}$ $(d \ll \mathcal{F})$ as the nonlinear projection operator, which maps the feature of the LR patch to an RKHS. The kernel function $k(\boldsymbol{l}_i, \boldsymbol{l}_j) = \phi(\boldsymbol{l}_i)^T \phi(\boldsymbol{l}_j)$ is employed to measure the nonlinear similarity between $\boldsymbol{l}_i$ and $\boldsymbol{l}_j$ in the projected high-dimensional space. We utilize a Gaussian kernel in this paper, which can be represented as $k(\boldsymbol{l}_i, \boldsymbol{l}_j) = exp(-\left\| \boldsymbol{l}_i - \boldsymbol{l}_j \right\|_2^2 / \upsilon)$. By using the operator ϕ, the corresponding LR training set can be represented in the RKHS as:

$$\begin{aligned} &\boldsymbol{l}_p^l \rightarrow \phi(\boldsymbol{l}_p^l) \\ &\mathbf{L}_p \rightarrow \phi(\mathbf{L}_p) = \left[\phi(\boldsymbol{l}_p^1), \ldots, \phi(\boldsymbol{l}_p^N)\right] \end{aligned} \tag{6}$$

Then, we utilize the mapped features of RKHS to substitute the LR features for reconstructing the HR patches. To preserve the reconstruction consistency between the ith LR patch and the corresponding HR patch, we can minimize the following local optimization function:

$$D_i(\mathbf{A}_p^\phi) = \left\| \boldsymbol{h}_p^i - \mathbf{A}^\phi \phi(\boldsymbol{l}_p^i) \right\|_2^2 \tag{7}$$

where $\mathbf{A}^\phi$ represents the kernel regression projection.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMM.2019.2898752, IEEE Transactions on Multimedia

Considering all the samples in the training set, we can summarize the local optimizations as a global objective function:

$$D(\mathbf{A}_p^\phi) = \left\|\mathbf{H}_p - \mathbf{A}_p^\phi \phi(\mathbf{L}_p)\right\|_F^2 + \lambda \left\|\mathbf{A}_p^\phi\right\|_F^2 \tag{8}$$

where the additional constraint is the regularization term, and the regularization parameter λ balances the above two terms.

The objective function (8) can be rewritten in terms of traces as:

$$\begin{aligned} D(\mathbf{A}_p^\phi) = & tr\left\{(\mathbf{H}_p - \mathbf{A}_p^\phi \phi(\mathbf{L}_p))(\mathbf{H}_p - \mathbf{A}_p^\phi \phi(\mathbf{L}_p))^T\right\} \\ & + \lambda tr\left\{\mathbf{A}_p^\phi (\mathbf{A}_p^\phi)^T\right\} \end{aligned} \tag{9}$$

B. Preserving the intrinsic geometric structure

As shown in Fig. 1, it is necessary to preserve the intrinsic distances between each given sample and its neighbors in the HR space after the reconstruction procedure, which means that the distance between the reconstruction results of sample $\boldsymbol{l}_p^i$ and its neighbor $\boldsymbol{l}_p^{i_k}$ should be close to the original distance between $\boldsymbol{h}_p^i$ and $\boldsymbol{h}_p^{i_k}$. For each training sample, we select K nearest neighbors to measure the intrinsic geometric structure. To represent the optimization function of the ith sample in mathematical formulation, we expect to minimize:

$$J_i(\mathbf{A}^\phi) = \sum_{k=1}^{K} \left\|(\boldsymbol{h}_p^i - \boldsymbol{h}_p^{i_k}) - (\mathbf{A}^\phi \phi(\boldsymbol{l}_p^i) - \mathbf{A}^\phi \phi(\boldsymbol{l}_p^{i_k}))\right\|_2^2 \omega_{i,i_k} \tag{10}$$

where the weight ω_{i,i^k} is utilized to measure the similarity between each HR patch $\boldsymbol{h}_p^i$ and its neighbor $\boldsymbol{h}_p^{i_k}$. It is defined as $\omega_{i,i_k} = \exp\left(-\left\|\boldsymbol{h}_p^i - \boldsymbol{h}_p^{i_k}\right\|_2^2 / \sigma^2\right)$, where the parameter σ is set to 1 in this paper.

Define the index set for neighbors of the ith data sample as $Q_i = \{i, i_1, \ldots, i_K\}$. The corresponding local data sets $\mathbf{H}_p^i = \left[\boldsymbol{h}_p^i, \boldsymbol{h}_p^{i_1} \ldots, \boldsymbol{h}_p^{i_K}\right]$ and $\mathbf{L}_p^i = \left[\boldsymbol{l}_p^i, \boldsymbol{l}_p^{i_1} \ldots, \boldsymbol{l}_p^{i_K}\right]$ represent the neighbors of the ith training sample in the HR and LR space, respectively. Equation (10) can be rewritten as:

$$\begin{aligned} & J_i(\mathbf{A}_p^\phi) \\ & = tr\left\{(\mathbf{H}_p^i \mathbf{S}_p^i - \mathbf{A}_p^\phi \phi(\mathbf{L}_p^i)\mathbf{S}_p^i)\mathbf{\Lambda}_p^i(\mathbf{H}_p^i \mathbf{S}_p^i - \mathbf{A}_p^\phi \phi(\mathbf{L}_p^i)\mathbf{S}_p^i)^T\right\} \\ & = tr\left\{(\mathbf{H}_p^i - \mathbf{A}_p^\phi \phi(\mathbf{L}_p^i))\mathbf{S}_p^i \mathbf{\Lambda}_p^i (\mathbf{S}_p^i)^T (\mathbf{H}_p^i - \mathbf{A}_p^\phi \phi(\mathbf{L}_p^i))^T\right\} \\ & = tr\left\{(\mathbf{H}_p^i - \mathbf{A}_p^\phi \phi(\mathbf{L}_p^i))\mathbf{\Sigma}_p^i (\mathbf{H}_p^i - \mathbf{A}_p^\phi \phi(\mathbf{L}_p^i))^T\right\} \end{aligned} \tag{11}$$

where the matrix $\mathbf{S}_p^i = \begin{bmatrix} \boldsymbol{e}_i^T \\ -\mathbf{I}_i \end{bmatrix} \in \mathcal{R}^{(K+1)\times K}$, $\boldsymbol{e}_i = [1, \ldots, 1]^T \in \mathcal{R}^K$, $\mathbf{I}_i$ is the $K \times K$ identity matrix, $\mathbf{\Sigma}_p^i = \mathbf{S}_p^i \mathbf{\Lambda}_p^i (\mathbf{S}_p^i)^T$ describes the intrinsic geometric structure, $\mathbf{\Lambda}_p^i = diag([\omega_{i,i_1}, \ldots, \omega_{i,i_K}])$ contains all the weighting coefficients on the diagonal, $diag(\cdot)$ is the diagonalization operator, and $tr(\cdot)$ is the trace operator.

Next, we incorporate the local optimizations of all the training samples for building the global alignment on the entire training set. Thus, we first unify all the local LR and HR data sets into a consistent coordinate system. The local coordinate $\mathbf{L}_p^i = \left[\boldsymbol{l}_p^i, \boldsymbol{l}_p^{i_1} \ldots, \boldsymbol{l}_p^{i_K}\right]$ and $\mathbf{H}_p^i = \left[\boldsymbol{h}_p^i, \boldsymbol{h}_p^{i_1} \ldots, \boldsymbol{h}_p^{i_K}\right]$ can be selected from the global coordinate $\mathbf{L}_p$ and $\mathbf{H}_p$ by utilizing a sample selection matrix $\mathbf{Z}_p^i \in \mathcal{R}^{N\times(K+1)}$:

$$\phi(\mathbf{L}_p^i) = \phi(\mathbf{L}_p)\mathbf{Z}_p^i, \quad \mathbf{H}_p^i = \mathbf{H}_p \mathbf{Z}_p^i \tag{12}$$

where the selection matrix $\mathbf{Z}_p^i$ can be described by:

$$(\mathbf{Z}_p^i)_{mn} = \begin{cases} 1 & if\ m = Q_i\{n\} \\ 0 & otherwise \end{cases} \tag{13}$$

Substituting (12) to (11), the local optimization of the ith sample can be rewritten as:

$$\begin{aligned} & J_i(\mathbf{A}_p^\phi) \\ & = tr\left\{(\mathbf{H}_p \mathbf{Z}_p^i - \mathbf{A}_p^\phi \phi(\mathbf{L}_p)\mathbf{Z}_p^i)\mathbf{\Sigma}_p^i (\mathbf{H}_p \mathbf{Z}_p^i - \mathbf{A}_p^\phi \phi(\mathbf{L}_p)\mathbf{Z}_p^i)^T\right\} \end{aligned} \tag{14}$$

By integrating all the local optimizations, the whole alignment for the entire training set can be represented as:

$$\begin{aligned} & J(\mathbf{A}_p^\phi) = \sum_i J_i(\mathbf{A}_p^\phi) \\ & = tr\left\{(\mathbf{H}_p - \mathbf{A}_p^\phi \phi(\mathbf{L}_p))(\sum_i \mathbf{Z}_p^i \mathbf{\Sigma}_p^i (\mathbf{Z}_p^i)^T)(\mathbf{H}_p - \mathbf{A}_p^\phi \phi(\mathbf{L}_p))^T\right\} \\ & = tr\left\{(\mathbf{H}_p - \mathbf{A}_p^\phi \phi(\mathbf{L}_p))\mathbf{\Theta}_p (\mathbf{H}_p - \mathbf{A}_p^\phi \phi(\mathbf{L}_p))^T\right\} \end{aligned} \tag{15}$$

where $\mathbf{\Theta}_p = \sum_i \mathbf{Z}_p^i \mathbf{\Sigma}_p^i (\mathbf{Z}_p^i)^T$ is the global alignment matrix.

C. Optimization

To maintain the reconstruction consistency and preserve the intrinsic geometric structure simultaneously, we combine the objective function in (9) and (15) to calculate the optimal kernel regression:

$$\mathbf{A}_p^\phi = \arg\min_{\mathbf{A}_p^\phi} F(\mathbf{A}_p^\phi; \mathbf{L}_p, \mathbf{H}_p) \tag{16}$$

It can be represented by:

$$\begin{aligned} & F(\mathbf{A}_p^\phi) \\ & = tr((\mathbf{H}_p - \mathbf{A}_p^\phi \phi(\mathbf{L}_p))(\mathbf{H}_p - \mathbf{A}_p^\phi \phi(\mathbf{L}_p))^T) + \lambda tr(\mathbf{A}_p^\phi (\mathbf{A}_p^\phi)^T) \\ & \quad + \mu tr((\mathbf{H}_p - \mathbf{A}_p^\phi \phi(\mathbf{L}_p))\mathbf{\Theta}_p (\mathbf{H}_p - \mathbf{A}_p^\phi \phi(\mathbf{L}_p))^T) \\ & = tr((\mathbf{H}_p - \mathbf{A}_p^\phi \phi(\mathbf{L}_p))(\mathbf{I} + \mu\mathbf{\Theta}_p)(\mathbf{H}_p - \mathbf{A}_p^\phi \phi(\mathbf{L}_p))^T) \\ & \quad + \lambda tr(\mathbf{A}_p^\phi (\mathbf{A}_p^\phi)^T) \end{aligned} \tag{17}$$

where μ is the regularization parameter to balance the two functions.

According to the kernel theory, the kernel regression $\mathbf{A}_p^\phi$ can be further described by the linear combination of $\phi(\mathbf{L}_p)$:

$$\mathbf{A}_p^\phi = \mathbf{C}_p \phi(\mathbf{L}_p)^T \tag{18}$$

where $\mathbf{C}_p \in \mathcal{R}^{m\times N}$ is the projection matrix. Thus, the optimization in (17) can be converted to calculate the optimal matrix $\mathbf{C}_p$. Define $\mathbf{K}_p = \phi(\mathbf{L}_p)^T \phi(\mathbf{L}_p)$ to be the kernel matrix in the RKHS. It can be reformulated by:

$$\begin{aligned} F(\mathbf{C}_p) = & tr((\mathbf{H}_p - \mathbf{C}_p \mathbf{K}_p)(\mathbf{I} + \mu\mathbf{\Theta}_p)(\mathbf{H}_p - \mathbf{C}_p \mathbf{K}_p)^T) \\ & + \lambda tr(\mathbf{C}_p \mathbf{K}_p \mathbf{C}_p^T) \end{aligned} \tag{19}$$

We calculate the gradient of (19) with respect to $\mathbf{C}_p$, and set it as $\mathbf{0}$:

$$-(\mathbf{H}_p - \mathbf{C}_p \mathbf{K}_p)(\mathbf{I} + \mu\mathbf{\Theta}_p)\mathbf{K}_p^T + \lambda \mathbf{C}_p \mathbf{K}_p = \mathbf{0} \tag{20}$$

According to the property of kernel function, we have $\mathbf{K}_p^T = \mathbf{K}_p$ in (20). Thus, the optimal projection matrix $\mathbf{C}_p$ can be deduced from (20) analytically:

$$\mathbf{C}_p = \mathbf{H}_p(\mathbf{I} + \mu\mathbf{\Theta}_p)[\mathbf{K}_p(\mathbf{I} + \mu\mathbf{\Theta}_p) + \lambda\mathbf{I}]^{-1} \tag{21}$$

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMM.2019.2898752, IEEE Transactions on Multimedia

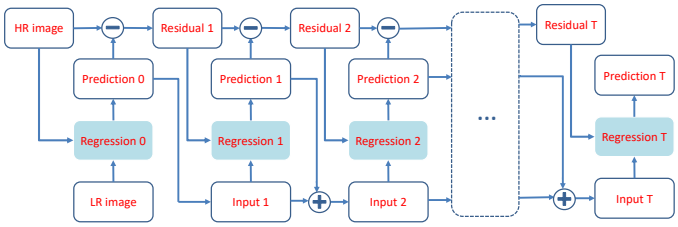


Fig. 2. The flowchart for the coarse-to-fine recursive regression structure.

Given an input testing LR patch $\boldsymbol{y}_p^L$, the corresponding HR patch $\boldsymbol{x}_p^H$ can be estimated by:

$$\begin{aligned}\boldsymbol{x}_p^H &= \mathbf{A}_p^{\phi}\phi(\boldsymbol{y}_p^L) \\ &= \mathbf{C}_p\phi(\mathbf{L}_p)^T\phi(\boldsymbol{y}_p^L) \\ &= \mathbf{H}_p(\mathbf{I}+\mu\mathbf{\Theta}_p)[\mathbf{K}_p(\mathbf{I}+\mu\mathbf{\Theta}_p)+\lambda\mathbf{I}]^{-1}\boldsymbol{\kappa}(\boldsymbol{y}_p^L)\end{aligned} \tag{22}$$

where the vector $\boldsymbol{\kappa}(\boldsymbol{y}_p^L) = [k(\boldsymbol{l}_p^1, \boldsymbol{y}_p^L), \ldots, k(\boldsymbol{l}_p^N, \boldsymbol{y}_p^L)]^T$.

D. The coarse-to-fine recursive regression structure

By optimizing (17), we can obtain a kernel regression to approximate the HR face patches by the input LR patches. However, the prediction of such a single regression model is not sufficiently precise, which causes the insufficiency of local details. Thus, we further propose a coarse-to-fine recursive regression structure to refine the high-frequency details. The training procedure is shown in Fig. 2. In iteration 0, we train the basic regression model $\mathbf{A}_p^{\phi 0}$ according to the input LR training set $\mathbf{L}_p^0$ and original HR training set $\mathbf{H}_p^0$. The regression model $\mathbf{A}_p^{\phi 0}$ can be employed to obtain the initial prediction of HR patches $\mathbf{P}_p^0 = \mathbf{A}_p^{\phi 0}\phi(\mathbf{L}_p^0)$ (i.e., Prediction 0 in Fig. 2). However, the high-frequency information may be lost in the initial prediction, which needs to be further generated by the following regressions. In each iteration t ($t = 1, 2, \ldots, T$), we continue to learn the corresponding regression model $\mathbf{A}_p^{\phi t}$ to approximate the residual high-frequency details $\mathbf{H}_p^t$ from the intermediate reconstructed patches $\mathbf{L}_p^t$ (i.e., Input t in Fig. 2). The ground truth of the residual high-frequency details is represented as $\mathbf{H}_p^t$ (i.e., Residual t in Fig. 2), while the predicted one is described as $\mathbf{P}_p^t$ (i.e., Prediction t in Fig. 2). It conducts a similar training strategy with (17) to learn the corresponding kernel regression function $\mathbf{A}_p^{\phi t}$. The iterative procedure can be formulated as:

$$\begin{aligned}\mathbf{L}_p^t &= \begin{cases}\mathbf{P}_p^{t-1}, t=1\\ \mathbf{L}_p^{t-1}+\mathbf{P}_p^{t-1}, t>2\end{cases} \\ \mathbf{H}_p^t &= \mathbf{H}_p^{t-1}-\mathbf{P}_p^{t-1} \\ \mathbf{A}_p^{\phi t} &= \arg\min_{\mathbf{A}_p^{\phi}} F(\mathbf{A}_p^{\phi t}; \mathbf{L}_p^t, \mathbf{H}_p^t) \\ \mathbf{P}_p^t &= \mathbf{A}_p^{\phi t}\phi(\mathbf{L}_p^t), t = 1, 2, \ldots, T\end{aligned} \tag{23}$$

To better enhance the local high-frequency information, we utilize the following four high-pass filters to extract the features from LR patches:

$$f_1 = [1, -1], f_2 = f_1^T, f_3 = [-1, 2, -1], f_4 = f_3^T. \tag{24}$$

These four descriptions are concatenated into one vector as the representation of each LR patch in $\mathbf{L}_p^0$. For the HR patches in $\mathbf{H}_p^0$, we subtract the mean value to reflect the local texture instead of absolute intensity. Assume that we have obtained all the kernel regression functions $\{\mathbf{A}_p^{\phi 0}, \mathbf{A}_p^{\phi 1}, \ldots, \mathbf{A}_p^{\phi T}\}$. For a testing LR patch $\boldsymbol{y}_p^L$, the hallucination task can be conducted by a recursive procedure:

$$\begin{aligned}\boldsymbol{x}_p^t &= \boldsymbol{x}_p^{t-1} + \mathbf{A}_p^{\phi(t-1)}\phi(\boldsymbol{x}_p^{t-1}) \\ \boldsymbol{x}_p^1 &= \mathbf{A}_p^{\phi 0}\phi(\boldsymbol{y}_p^L), t = 1, 2, \ldots, T\end{aligned} \tag{25}$$

Finally, we can reconstruct the hallucinated HR patch $\boldsymbol{x}_p^T$ by proceeding (25) T times. We also add the mean value $\bar{x}_p$ to $\boldsymbol{x}_p^T$ in order to recover the absolute intensity, where the mean value is estimated by the input LR patch. The whole HR face image is produced by integrating all the HR patches according to their original positions. The pixel values in the overlapped regions are calculated by averaging the pixels of adjacent reconstructed patches.

IV. Experimental Results

To evaluate the performance of the proposed method, we perform the experiments on the CAS-PEAL [51] and FERET [52] face databases for comparison. We utilize a subset of the CAS-PEAL database which includes 1040 frontal face images to conduct the experiment. We randomly select 40 images to serve as the testing set, while considering the remaining 1000 images as the training set. For the experiments on the FERET database, we utilize 600 face images from three subsets (i.e., ba, bj and bk) which were captured by 200 various individuals. The subset ba includes face images with normal expressions. The subset bj contains face images with expression variations. The subset bk consists of face images under different illumination conditions. We employ 30 images from 10 individuals as the testing set, while utilizing the other 570 images as the training set. To conduct the experiments, we align all the face images by the position of two eyes and crop them to 128×128 pixels. We then blur the HR images by utilizing a 7×7 Gaussian filter with a standard deviation of 0.85 and generate LR images of 32×32 pixels by a down-sampling process. For experimental convenience, we convert the intensity values of the face images to the range from 0 (black) to 1 (white). To demonstrate the effectiveness of the proposed method, we compare the experimental results with several state-of-the-art baselines (e.g., LSR [25], LcR [30], WASR [28], SRLSP [22]) and three deep learning based methods (e.g., VDSR [35], LCGE [39] and WaveSR [42]).

A. Parameter settings

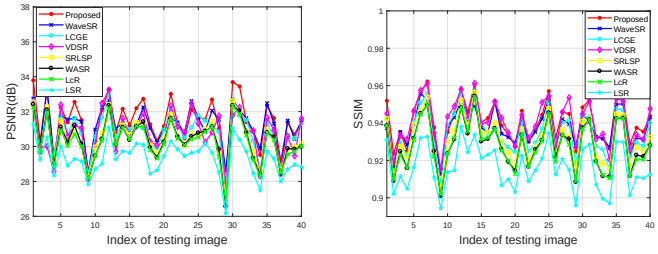
There are some parameters which require to be determined before conducting the experiments. We utilize a Gaussian kernel to project the feature of the LR patch into RKHS, while the parameter υ in the Gaussian function is set to 3. In Section III-B, the number of neighboring samples is set to $K = 200$ for preserving the intrinsic geometric structure. To optimize Eq. (17), we fix the regularization parameters λ and μ to 0.5 and $1/K$, respectively. The maximum number of iterations is set to three.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMM.2019.2898752, IEEE Transactions on Multimedia

Fig. 3. The hallucinated results on the CAS-PEAL database. (Please zoom in the results for better comparison. The main distinctions are highlighted in the rectangle.) (a) LR image. (b) LSR [25]. (c) LcR [30]. (d) WASR [28]. (e) SRLSP [22]. (f) VDSR [35]. (g) LCGE [39]. (h) WaveSR [42]. (i) The proposed method. (j) Original HR image.

B. Experimental results on the CAS-PEAL database

To show the effectiveness of the proposed approach, we evaluate the performance with seven state-of-the-art methods in this subsection, which include LSR [25], LcR [30], WASR [28], SRLSP [22], VDSR [35], LCGE [39] and WaveSR [42]. We present some representative reconstructed HR images in Fig. 3. As shown in the figure, the results of LSR [25] suffer from smoothing artifacts around the regions of the nose, eyes and mouth, while the artificial effects also occur near the contour of the face image. LcR [30] has the ability to enhance local details around the eyes and nostrils, but the ringing artifacts and noises are not well suppressed on the face image. WASR [28] improves artificial effects around the contour of the face image. However, the textures around the eyes and mouth seem to be blurred after reconstruction. SRLSP [22] reduces the jaggy and noisy effects on the hallucinated results, but the aliasing effects appear on the regions with high-frequency textures. VDSR [35] produces clear textures on the cheek and the contour of face image. However, it sometimes suffers from artificial details especially on the region around the eyes. LCGE [39] reconstructs fine results on most human face regions. However, it also causes noticeable artificial effects around the area of the eyes. For example, it fails to reconstruct the shape of two-layer eyelids. WaveSR [42] further improves the visual quality of the hallucinated images, which is helpful to produce facial components with more plausible textures. The proposed approach successfully suppresses the noise and eliminates the artifacts of ringing effects on the reconstructed face image. Furthermore, it also produces more reasonable results on the regions with vivid details (e.g., eyeballs, nose and mouth). Compared with other methods, the proposed method also presents the advantage of recovering subtle details on the facial components, such as eyelid, pouch, nose and mouth.



(a) (b)

Fig. 4. The quantitative indexes on the CAS-PEAL database. (a) PSNR. (b) SSIM.

In Fig. 4, we also present the comparison of the PSNR and SSIM [53] indexes on the hallucinated face images for evaluating the above algorithms. As shown in the figure, the proposed method obtains the highest quantitative results for most testing images. In Table I, we also present the average PSNR and

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMM.2019.2898752, IEEE Transactions on Multimedia

TABLE I
THE AVERAGE QUANTITATIVE INDEXES ON THE CAS-PEAL DATABASE.

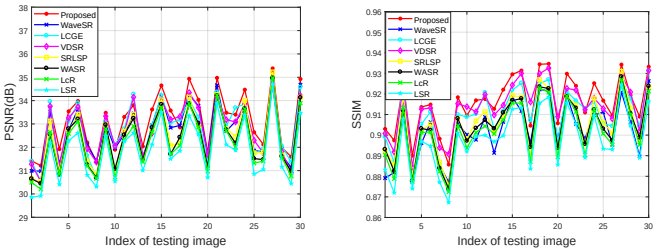
	LSR	LcR	WASR	SRLSP	VDSR	LCGE	WaveSR	Proposed
PSNR(dB)	29.30	30.26	30.36	30.49	30.79	30.85	31.23	**31.43**
SSIM	0.9160	0.9279	0.9283	0.9317	0.9408	0.9348	0.9376	**0.9412**

TABLE II
THE AVERAGE QUANTITATIVE INDEXES ON THE FERET DATABASE.

	LSR	LcR	WASR	SRLSP	VDSR	LCGE	WaveSR	Proposed
PSNR (dB)	31.86	32.21	32.33	32.44	32.79	32.78	32.72	**33.15**
SSIM	0.8982	0.9021	0.9043	0.9056	0.9139	0.9097	0.9027	**0.9164**

SSIM indexes for different methods. The above experimental results demonstrate the advantage of the proposed algorithm according to both qualitative and quantitative comparisons.

C. Experimental results on the FERET database



(a) (b)

Fig. 6. The quantitative indexes on the FERET database. (a) PSNR. (b) SSIM.

In this subsection, we further demonstrate the advantages of the proposed method by performing the experiments on the FERET database. The face images in the FERET database are captured under various illumination conditions and facial expressions. The experiments on such a database can evaluate the flexibility of different methods in alternative environments. We present some hallucinated results in Fig. 5. The proposed algorithm obtains fine textures on the cheek region and reconstructs vivid details on regions around the eyes and mouth. Meanwhile, the compared methods seem to produce some artificial effects on the reconstructed face image. For example, the textures of mouth suffer from blurring effects, while some unreal details also appear near the area of the eyes. Fig. 6 shows the corresponding quantitative indexes for comparing the results of different methods. The average quantitative indexes are also presented in Table II. Compared with the second best quantitative indexes, the result of the proposed algorithm increases 0.36dB in PSNR and 0.0025 in SSIM. The above compared results indicate that the proposed approach consistently achieves superior performance under various environments.

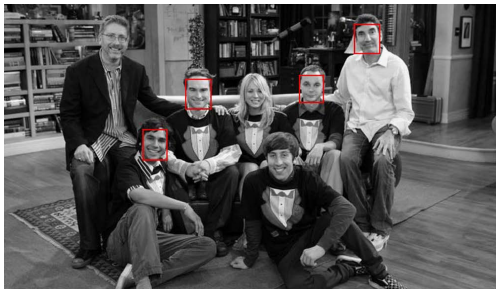


Fig. 7. One real-world image captured in the real environment.

D. Experiments on real-world images

To further evaluate the performance of the proposed method on a real-world image, we collect one image from the Internet which was captured in the real environment. The selected image is shown in Fig. 7. We conduct the face hallucination procedure on the highlighted human faces. In the experiment, we manually align the LR face images and standardize the size to 32×32 pixels, while the FERET database is selected as the training image set to hallucinate the face images. The experimental results are shown in Fig. 8 for comparison. The image captured from the real environment contains more variations, which considerably influences the performance of LSR [25], LcR [30] and WASR [28]. For these algorithms, the hallucinated results are also obscured around the eye region (especially the eyelid), which seems to be affected by aliasing artifacts. The other three methods [22], [35] and [39] have the advantage of alleviating the noisy artifacts, but the reconstructed face images suffer from apparent blurring effects or artifacts around the eyeballs. WaveSR [42] has the ability to reconstruct reasonable facial components, whereas the noise seems to be magnified on the face images. The proposed method successfully suppresses the noise and reconstructs more suitable details on the regions with complex textures, which is able to produce acceptable results.

E. Influence of training set size

To study the influence of the training set size in the face hallucination problem, we further perform experiments on the CAS-PEAL face database by varying the number of training images from 1000 to 200 with step 200. We also fix the testing samples in the previous experiments to conduct fair comparisons. Then, we perform the experiments on various training sets and evaluate the impact of the training set size. Fig. 9 shows some experimental results with different training set sizes. The average quantitative indexes for the 40 testing images are presented in Table III. When the number of training images is larger than 600, the visual quality of the reconstructed images varies very slightly with changes in the training set size. Similar conclusions can also be drawn from Table III. When the number of training images decreases to 200, the visual quality of the hallucinated HR images suffers from degradation near the contour of face images. However, it consistently suppresses the noisy artifacts and preserves vivid high-frequency details, especially on the region of the facial components (e.g., nose, eyes and mouth).

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMM.2019.2898752, IEEE Transactions on Multimedia

Fig. 5. The hallucinated results on the FERET database. (Please zoom in the results for better comparison. The main distinctions are highlighted in the rectangle.) (a) LR image. (b) LSR [25]. (c) LcR [30]. (d) WASR [28]. (e) SRLSP [22]. (f) VDSR [35]. (g) LCGE [39]. (h) WaveSR [42]. (i) The proposed method. (j) Original HR image.

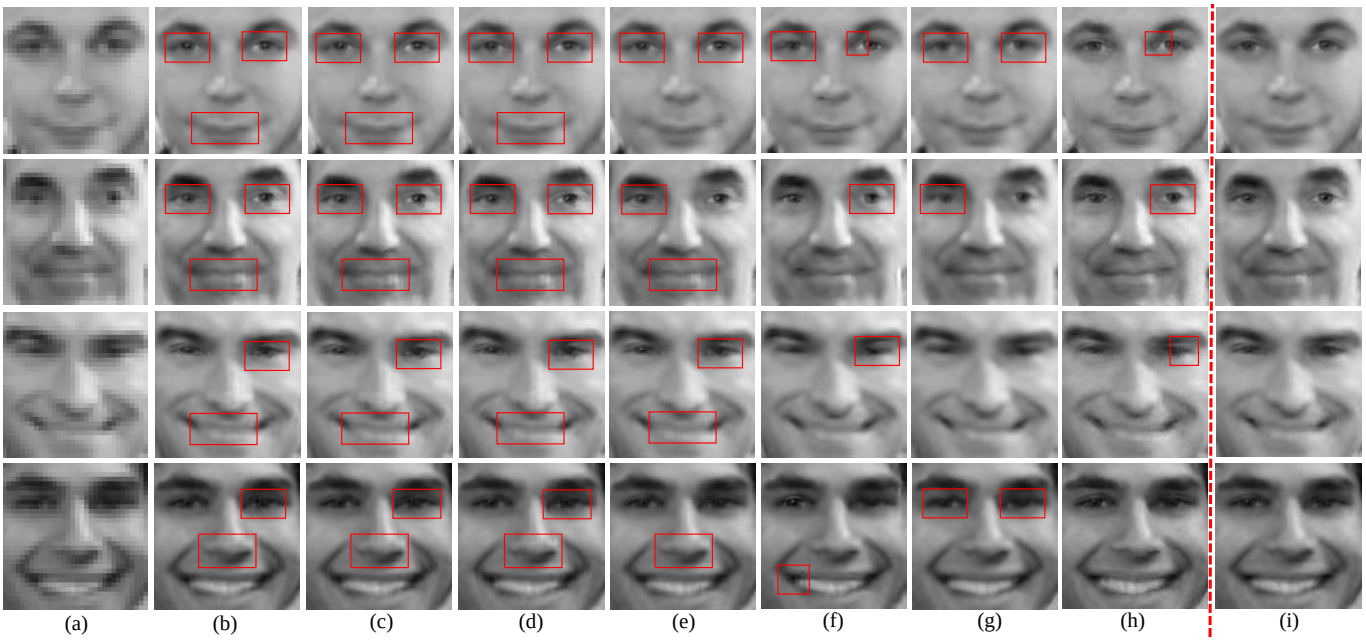


Fig. 8. The hallucinated results for the real-world images. (Please zoom in the results for better comparison. The main distinctions are highlighted in the rectangle.)(a) LR image. (b) LSR [25]. (c) LcR [30]. (d) WASR [28]. (e) SRLSP [22]. (f) VDSR [35]. (g) LCGE [39]. (h) WaveSR [42]. (i) The proposed method.

F. Effectiveness of using kernel regression

In this subsection, we experimentally show the effectiveness of using the kernel mapping function to improve the face hallucination results under the proposed framework. To achieve this objective, we utilize linear regression instead of

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMM.2019.2898752, IEEE Transactions on Multimedia

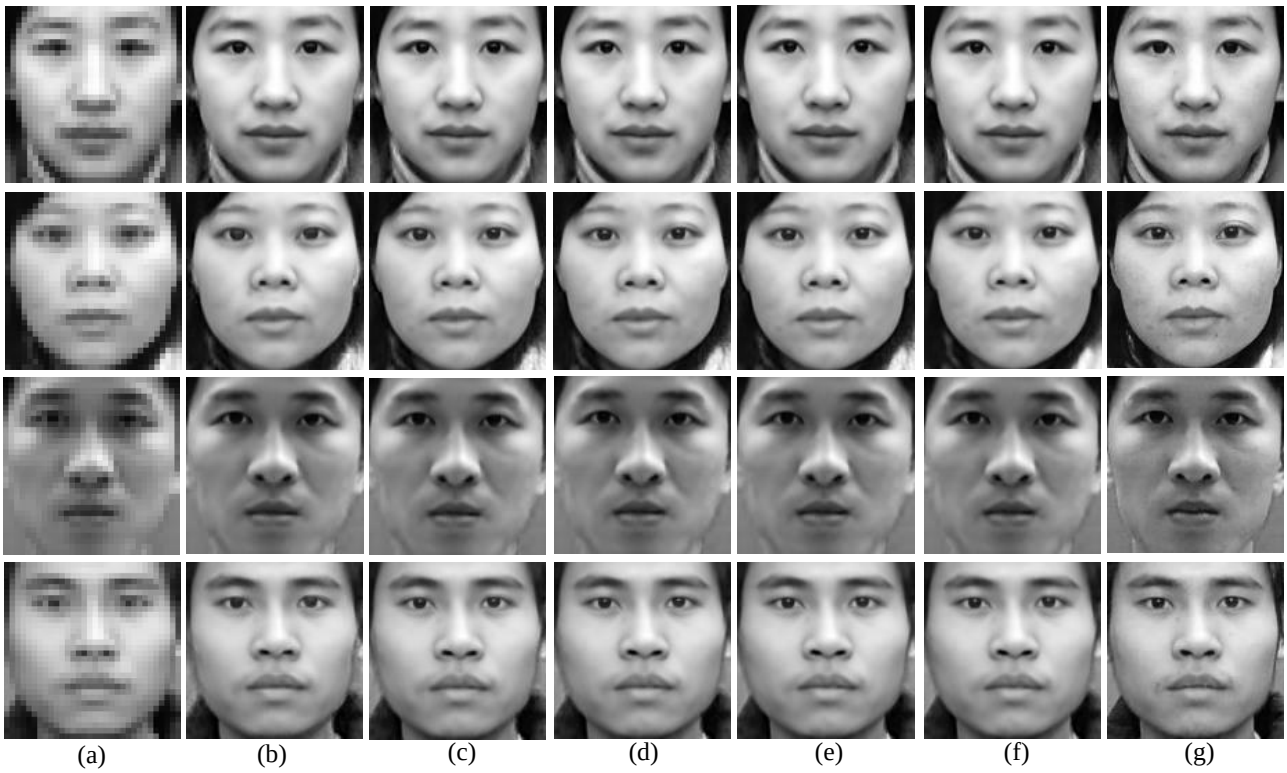


Fig. 9. The hallucinated results on the CAS-PEAL database with different training set sizes. (a) LR image. (b) With 200 samples. (c) With 400 samples. (d) With 600 samples. (e) With 800 samples. (f) With 1000 samples. (g) Original HR image.

TABLE III
THE AVERAGE QUANTITATIVE INDEXES FOR VARIOUS TRAINING SET SIZES.

Size	200	400	600	800	1000
PSNR (dB)	30.01	30.83	31.15	31.32	31.43
SSIM	0.9254	0.9357	0.9390	0.9407	0.9412

kernel regression to conduct the face hallucination procedure. Then, we compare the linear regression based results with the experimental results of the proposed method. Specifically, we can rewrite (17) to represent the hallucination model with linear regression:

$$F(\mathbf{A}_p)=tr((\mathbf{H}_p-\mathbf{A}_p\mathbf{L}_p)(\mathbf{I}+\mu\mathbf{\Theta}_p)(\mathbf{H}_p-\mathbf{A}_p\mathbf{L}_p)^T) + \lambda tr(\mathbf{A}_p\mathbf{A}_p^T) \tag{26}$$

where $\mathbf{A}_p$ is the linear mapping function that corresponds to position p. We calculate the gradient of (26) and set it as $\mathbf{0}$:

$$-(\mathbf{H}_p-\mathbf{A}_p\mathbf{L}_p)(\mathbf{I}+\mu\mathbf{\Theta}_p)\mathbf{L}_p^T+\lambda\mathbf{A}_p=\mathbf{0} \tag{27}$$

Thus, the optimal linear mapping function $\mathbf{A}_p$ can be deduced from (27):

$$\mathbf{A}_p=\mathbf{H}_p(\mathbf{I}+\mu\mathbf{\Theta}_p)\mathbf{L}_p^T\left[\mathbf{L}_p(\mathbf{I}+\mu\mathbf{\Theta}_p)\mathbf{L}_p^T+\lambda\mathbf{I}\right]^{-1} \tag{28}$$

Given an input testing LR patch $\boldsymbol{y}_p^L$, the output HR patch $\boldsymbol{x}_p^H$ can be estimated by:

$$\begin{aligned}\boldsymbol{x}_p^H&=\mathbf{A}_p\boldsymbol{y}_p^L\\&=\mathbf{H}_p(\mathbf{I}+\mu\mathbf{\Theta}_p)\mathbf{L}_p^T\left[\mathbf{L}_p(\mathbf{I}+\mu\mathbf{\Theta}_p)\mathbf{L}_p^T+\lambda\mathbf{I}\right]^{-1}\boldsymbol{y}_p^L\end{aligned} \tag{29}$$

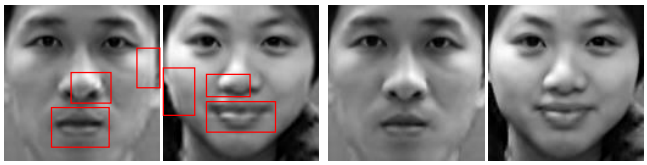
which can be utilized to substitute (22) for the experiment with linear regression.

We conduct the experiment on the CAS-PEAL database to verify the effectiveness of using kernel regression in the reconstruction. We also fix the 40 testing images in the experiment for comparing the visual quality and quantitative indicators. Some representative hallucinated results are presented in Fig. 10. The proposed algorithm obtains superior hallucinated results especially on the regions of the nose, mouth and cheek. The quantitative indicators for all the testing images are presented in Fig. 11. It is apparent that the reconstruction by kernel regression produces better results than the algorithm performed by linear regression. The average PSNR and SSIM values of all the testing images are 29.82dB and 0.9250 for the face hallucination algorithm with linear regression. By taking advantage of kernel regression, the proposed method achieves an improvement of 1.61dB in PSNR and 0.0162 in SSIM. The experimental results demonstrate that the utilization of kernel regression is effective to obtain superior HR images for the proposed face hallucination framework.

G. Effectiveness of the geometric structure constraint

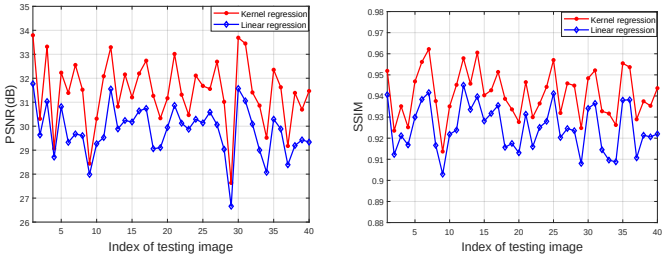
We further discuss the contribution of the intrinsic geometric structure to show the effectiveness. To achieve this objective, we eliminate the corresponding intrinsic geometric structure constraint in (17) by setting $\mu=0$, which presents the performance of the algorithm without the intrinsic geometric structure for comparison. Then, we compare the hallucinated results with the proposed method to show the advantages

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMM.2019.2898752, IEEE Transactions on Multimedia



(a) (b)

Fig. 10. Comparison for the performance of utilizing different mapping function. (a) Linear regression. (b) Kernel regression.



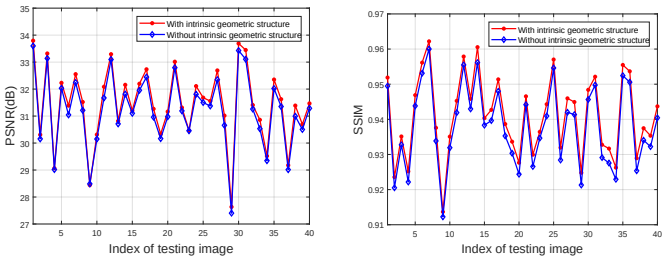
(a) (b)

Fig. 11. The quantitative indicators for comparison. (a) PSNR. (b) SSIM.

(a) (b)

Fig. 12. Comparison for utilizing the intrinsic geometric structure constraint. (a) Without intrinsic geometric structure. (b) With intrinsic geometric structure.

of intrinsic geometric structure constraint. The experiments are also conducted on the CAS-PEAL database, while some representative hallucinated results are shown in Fig. 12. The corresponding PSNR and SSIM values are also presented in Fig. 13. The algorithm without intrinsic geometric structure constraint achieves the average quantitative indicators of 31.21dB in PSNR and 0.9379 in SSIM, which are inferior to the proposed approach by 0.22dB in PSNR and 0.0033 in SSIM. According to the experimental results, we can conclude that the intrinsic geometric structure constraint truly benefits the reconstruction results in the proposed face hallucination approach.

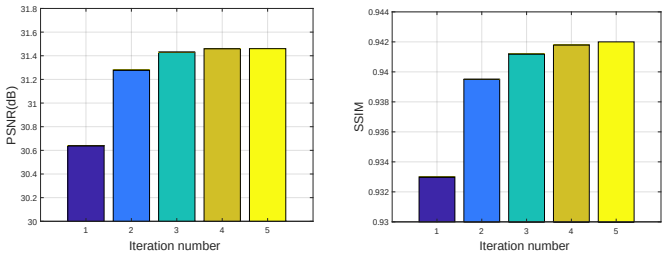


(a) (b)

Fig. 13. The quantitative indicators for comparison. (a) PSNR. (b) SSIM.

H. Influence of the recursive regression structure

To evaluate the influence of the recursive regression structure in the proposed algorithm, we present the average PSNR and SSIM values along with the variation of iterations on the CAS-PEAL database. The results are shown in Fig. 14. The proposed algorithm obtains better performance according to the increase in iterations, which indicates that the residual high-frequency details are well compensated by the recursive regression structure. The proposed method achieves 30.64dB in PSNR and 0.9330 in SSIM with only one regression step, while the quantitative indicators increase to 31.46 dB in PSNR and 0.9420 in SSIM after five regression steps. Accordingly, a higher computational cost is required together with an increase of regression steps. In practice, the proposed method can roughly achieve stable results after only three regression steps. Thus, we fix the maximum number of recursive steps to three in the proposed approach. It is also possible to adjust the number of recursive steps in the perspective of the balance between quality and efficiency.



(a) (b)

Fig. 14. The PSNR and SSIM indicators with different numbers of iterations. (a) PSNR. (b) SSIM.

I. Computational Cost

To investigate the capability of various methods, we also compare the average computational time of the proposed algorithm with other state-of-the-art methods on the CAS-PEAL database. Specifically, we perform LSR [25], LcR [30], WASR [28], SRLSP [22] and the proposed method using MATLAB 2017a on a computer of 8G memory and 3.2 GHz CPU. VDSR [35] and WaveSR [42] are conducted on a single Tesla K80 GPU by Caffe and PyTorch, respectively. LCGE [39] is implemented by Caffe on a Tesla K80 GPU, while the local detail enhancement process is conducted on CPU using C++. The average computational time of various methods is presented in Table IV. Compared with other methods, the proposed algorithm produces superior results with reasonable computational cost because the proposed method only requires several regression procedures for generating the final HR image. The computational cost can be further reduced by decreasing the number of iterations, which also sightly degrades the quality of the reconstructed HR image.

J. Performance of the face recognition task

We further conduct experiments on the low-resolution face recognition task to verify the effectiveness of face hallucination algorithms for improving performance. The evaluation

TABLE IV
COMPARISON FOR THE AVERAGE RUNNING TIME OF VARIOUS FACE HALLUCINATION ALGORITHMS. NOTICE THAT THE DEEP LEARNING BASED APPROACHES ARE IMPLEMENTED ON GPU.

LSR	LcR	WASR	SRLSP	VDSR	LCGE	WaveSR	Proposed
10.2s	10.9s	24.1s	9.1s	0.6s	68.9s	3.3s	3.6s

TABLE V
THE RECOGNITION ACCURACY FOR DIFFERENT FACE HALLUCINATION ALGORITHMS

LSR	LcR	WASR	SRLSP	VDSR
96.71%	97.58%	97.84%	97.75%	98.87%
LCGE	WaveSR	Proposed	HR	
97.75%	99.57%	**99.65%**	99.91%	

is performed on a subset of the Multi-PIE face database [54], which includes frontal face images under 20 different illumination conditions (Session 04, Camera: 05_1, Recording No.01). For each subject, we utilize five face images with various illumination to constitute the testing set, while employing the remaining face images to serve as the training set. In the experiment, we normalize the HR face images to the resolution of 64×64 pixels, while obtaining the LR images of 16×16 pixels by down-sampling. By using various face hallucination methods, we first magnify the LR testing images to the resolution of 64×64 pixels. Then, the hallucinated face images are utilized to evaluate the face recognition performance at the HR level. We employ ResNet [55] to conduct the face recognition procedure. Before entering the first residual block, a convolutional layer with kernel size 7×7, stride of 2, and 64 output channels is performed on the input images. Then, we utilize a group of residual blocks in the network, which has the same structure as ResNet18. At the end of the last residual block, the network is followed by a global average pooling, a fully-connected layer and softmax. We employ PyTorch for implementation and use SGD as the optimizer. The batch size is set to 128. A weight decay of 0.0005 and a momentum of 0.9 are utilized to train the network. The learning rate starts from 0.1 and is divided by 10 after 50 epochs until the loss is steady. The experimental results for different hallucination approaches are presented in Table V. We also supply the recognition rate on the original HR testing images for comparison. The proposed method obtains the accuracy of 99.65%, which is the highest recognition rate among the compared algorithms. The above experiments demonstrate that the proposed approach can effectively reconstruct plausible high-frequency information and benefit the face recognition task.

V. CONCLUSION

In this paper, we propose a novel face hallucination algorithm that reconstructs the HR face image with a series of adaptive regression mappings by a coarse-to-fine strategy. In the training phase of each regression, the proposed method takes advantage of kernel function to seek the nonlinear characteristic, while simultaneously considering the reconstruction consistency and preserving the intrinsic geometric structure between neighboring samples. The coarse-to-fine strategy is achieved by the recursive regression structure, which first generates an initial HR face image and then compensates for the residual reconstruction error by an iterative procedure. In the testing phase, the final hallucinated face image can be simply obtained by the projection of regression mappings, which induces the efficiency of the proposed method. The proposed approach is rather fast yet effective for reconstructing HR face images. Experimental results demonstrate that the proposed algorithm produces superior performance with reasonable computational time when compared with the state-of-the-art algorithms.

ACKNOWLEDGEMENT

The authors wish to acknowledge CSC-IT Center for Science, Finland, for computational resources.

REFERENCES

[1] N. Wang, D. Tao, X. Gao, X. Li, and J. Li, "A comprehensive survey to face hallucination," *International Journal of Computer Vision*, vol. 106, no. 1, pp. 9–30, 2014.
[2] X. Zhang and X. Wu, "Image interpolation by adaptive 2-D autoregressive modeling and soft-decision estimation," *IEEE Trans. on Image Processing*, vol. 17, no. 6, pp. 887–896, 2008.
[3] K. Zhang, X. Gao, D. Tao, and X. Li, "Single image super-resolution with non-local means and steering kernel regression," *IEEE Transactions on Image Processing*, vol. 21, no. 11, pp. 4544–4556, 2012.
[4] J. Sun, Z. Xu, and H.-Y. Shum, "Gradient profile prior and its applications in image super-resolution and enhancement," *IEEE Trans. on Image Processing*, vol. 20, no. 6, pp. 1529–1542, 2011.
[5] W. Dong, L. Zhang, G. Shi, and X. Li, "Nonlocally centralized sparse representation for image restoration," *IEEE Transactions on Image Processing*, vol. 22, no. 4, pp. 1620–1630, 2013.
[6] W. Dong, F. Fu, G. Shi, X. Cao, J. Wu, G. Li, and X. Li, "Hyperspectral image super-resolution via non-negative structured sparse representation," *IEEE Transactions on Image Processing*, vol. 25, no. 5, pp. 2337–2352, 2016.
[7] J. Jiang, X. Ma, C. Chen, T. Lu, Z. Wang, and J. Ma, "Single image super-resolution via locally regularized anchored neighborhood regression and nonlocal means," *IEEE Transactions on Multimedia*, vol. 19, no. 1, pp. 15–26, 2017.
[8] H. Chen, X. He, L. Qing, and Q. Teng, "Single image super-resolution via adaptive transform-based nonlocal self-similarity modeling and learning-based gradient regularization," *IEEE Transactions on Multimedia*, vol. 19, no. 8, pp. 1702–1717, 2017.
[9] Y. Li, W. Dong, X. Xie, G. Shi, J. Wu, and X. Li, "Image super-resolution with parametric sparse model learning," *IEEE Transactions on Image Processing*, vol. 27, no. 9, pp. 4638–4650, 2018.
[10] R. Timofte, V. De Smet, and L. Van Gool, "A+: Adjusted anchored neighborhood regression for fast super-resolution," in *Asian Conference on Computer Vision*. Springer, 2014, pp. 111–126.
[11] S. Baker and T. Kanade, "Hallucinating faces," in *Proceedings of the IEEE International Conference on Automatic Face and Gesture Recognition*, 2000, pp. 83–88.
[12] H. Chang, D. Yeung, and Y. Xiong, "Super-resolution through neighbor embedding," in *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition*, 2004, pp. 275–282.
[13] X. Wang and X. Tang, "Hallucinating face by eigentransformation," *IEEE Trans. on Systems, Man, and Cybernetics, Part C: Applications and Reviews*, vol. 35, no. 3, pp. 425–434, 2005.
[14] J. S. Park and S. W. Lee, "An example-based face hallucination method for single-frame, low-resolution facial images," *IEEE Trans. on Image Processing*, vol. 17, no. 10, pp. 1806–1816, 2008.
[15] C. Liu, H.-Y. Shum, and W. T. Freeman, "Face hallucination: Theory and practice," *International Journal of Computer Vision*, vol. 75, no. 1, pp. 115–134, 2007.
[16] Y. Zhuang, J. Zhang, and F. Wu, "Hallucinating faces: LPH super-resolution and neighbor reconstruction for residue compensation," *Pattern Recognition*, vol. 40, no. 11, pp. 3178–3194, 2007.
[17] K. Jia and S. Gong, "Generalized face super-resolution," *IEEE Transactions on Image Processing*, vol. 17, no. 6, pp. 873–886, 2008.

[18] H. Huang, H. He, X. Fan, and J. Zhang, "Super-resolution of human face image using canonical correlation analysis," *Pattern Recognition*, vol. 43, no. 7, pp. 2532–2543, 2010.

[19] J. Yang, J. Wright, T. Huang, and Y. Ma, "Image super-resolution via sparse representation," *IEEE Trans. on Image Processing*, vol. 19, no. 11, pp. 2861–2873, 2010.

[20] L. An and B. Bhanu, "Face image super-resolution using 2d cca," *Signal Processing*, vol. 103, pp. 184–194, 2014.

[21] H. Huang and N. Wu, "Fast facial image super-resolution via local linear transformations for resource-limited applications," *IEEE Trans. on Circuits and Systems for Video Technology*, vol. 21, no. 10, pp. 1363–1377, 2011.

[22] J. Jiang, C. Chen, J. Ma, Z. Wang, Z. Wang, and R. Hu, "SRLSP: A face image super-resolution algorithm using smooth regression with local structure prior," *IEEE Transactions on Multimedia*, vol. 19, no. 1, pp. 27–40, 2017.

[23] W. Zhang and W.-K. Cham, "Hallucinating face in the DCT domain," *IEEE Trans. on Image Processing*, vol. 20, no. 10, pp. 2769–2779, 2011.

[24] B. Li, H. Chang, S. Shan, and X. Chen, "Aligning coupled manifolds for face hallucination," *IEEE Signal Processing Letters*, vol. 16, no. 11, pp. 957–960, 2009.

[25] X. Ma, J. Zhang, and C. Qi, "Hallucinating face by position-patch," *Pattern Recognition*, vol. 43, no. 6, pp. 2224–2236, 2010.

[26] X. Ma, H. Song, and X. Qian, "Robust framework of single-frame face superresolution across head pose, facial expression, and illumination variations," *IEEE Transactions on Human-Machine Systems*, vol. 45, no. 2, pp. 238–250, 2015.

[27] C. Jung, L. Jiao, B. Liu, and M. Gong, "Position-patch based face hallucination using convex optimization," *IEEE Signal Processing Letters*, vol. 18, no. 6, pp. 367–370, 2011.

[28] Z. Wang, R. Hu, S. Wang, and J. Jiang, "Face hallucination via weighted adaptive sparse regularization," *IEEE Transactions on Circuits and Systems for video Technology*, vol. 24, no. 5, pp. 802–813, 2014.

[29] L. Liu, C. P. Chen, S. Li, Y. Y. Tang, and L. Chen, "Robust face hallucination via locality-constrained bi-layer representation," *IEEE transactions on cybernetics*, vol. 48, no. 4, pp. 1189–1201, 2018.

[30] J. Jiang, R. Hu, Z. Wang, and Z. Han, "Noise robust face hallucination via locality-constrained representation," *IEEE Transactions on Multimedia*, vol. 16, no. 5, pp. 1268–1281, 2014.

[31] J. Shi and C. Qi, "Kernel-based face hallucination via dual regularization priors," *IEEE Signal Processing Letters*, vol. 22, no. 8, pp. 1189–1193, 2015.

[32] J. Shi, X. Liu, and C. Qi, "Global consistency, local sparsity and pixel correlation: A unified framework for face hallucination," *Pattern Recognition*, vol. 47, no. 11, pp. 3520–3534, 2014.

[33] X. Zeng, H. Huang, and C. Qi, "Expanding training data for facial image super-resolution," *IEEE transactions on cybernetics*, vol. 48, no. 2, pp. 716–729, 2018.

[34] C. Dong, C. C. Loy, K. He, and X. Tang, "Image super-resolution using deep convolutional networks," *IEEE transactions on pattern analysis and machine intelligence*, vol. 38, no. 2, pp. 295–307, 2016.

[35] J. Kim, J. Kwon Lee, and K. Mu Lee, "Accurate image super-resolution using very deep convolutional networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 1646–1654.

[36] X. Yu and F. Poriklii, "Ultra-resolving face images by discriminative generative networks." in *ECCV*, 2016, pp. 318–333.

[37] X. Yu and F. Porikli, "Hallucinating very low-resolution unaligned and noisy face images by transformative discriminative autoencoders," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 3760–3768.

[38] S. Zhu, S. Liu, C. C. Loy, and X. Tang, "Deep cascaded bi-network for face hallucination." in *ECCV*, 2016, pp. 614–630.

[39] Y. Song, J. Zhang, S. He, L. Bao, and Q. Yang, "Learning to hallucinate face images via component generation and enhancement," in *International Joint Conference on Artificial Intelligence (IJCAI)*, 2017, pp. 4537–4543.

[40] Q. Cao, L. Lin, Y. Shi, X. Liang, and G. Li, "Attention-aware face hallucination via deep reinforcement learning," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 690–698, 2017.

[41] Y. Chen, Y. Tai, X. Liu, C. Shen, and J. Yang, "FSRNet: End-to-end learning face super-resolution with facial priors," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 2492–2501.

[42] H. Huang, R. He, Z. Sun, and T. Tan, "Wavelet-srnet: A wavelet-based cnn for multi-scale face super resolution," in *Proceedings of the IEEE Conference on Computer Vision*, 2017, pp. 1689–1697.

[43] X. Yu, B. Fernando, B. Ghanem, F. Porikli, and R. Hartley, "Face super-resolution guided by facial component heatmaps," in *Proceedings of the European Conference on Computer Vision*, 2018, pp. 217–233.

[44] J. Shi, X. Liu, Y. Zong, C. Qi, and G. Zhao, "Hallucinating face image by regularization models in high-resolution feature space," *IEEE Transactions on Image Processing*, vol. 27, no. 6, pp. 2980–2995, 2018.

[45] J. Jiang, R. Hu, Z. Wang, Z. Han, and J. Ma, "Face super-resolution via multilayer locality-constrained iterative neighbor embedding and intermediate dictionary learning," *IEEE Transactions on Image Processing*, vol. 23, no. 10, pp. 4220–4231, 2014.

[46] Y. Hu, N. Wang, D. Tao, X. Gao, and X. Li, "SERF: a simple, effective, robust, and fast image super-resolver from cascaded linear regression," *IEEE Transactions on Image Processing*, vol. 25, no. 9, pp. 4091–4102, 2016.

[47] K. Zhang, D. Tao, X. Gao, X. Li, and J. Li, "Coarse-to-fine learning for single-image super-resolution," *IEEE transactions on neural networks and learning systems*, vol. 28, no. 5, pp. 1109–1122, 2017.

[48] H. Chen, X. He, L. Qing, Q. Teng, and C. Ren, "SGCRSR: Sequential gradient constrained regression for single image super-resolution," *Signal Processing: Image Communication*, vol. 66, pp. 1–18, 2018.

[49] K. Zhang, Z. Wang, J. Li, X. Gao, and Z. Xiong, "Learning recurrent residual regressors for single image super-resolution," *Signal Processing*, vol. 154, pp. 324–337, 2019.

[50] T. Hofmann, B. Scholkopf, and A. J. Smola, "Kernel methods in machine learning," *The Annals of Statistics*, pp. 1171–1220, 2008.

[51] W. Gao, B. Cao, S. Shan, X. Chen, D. Zhou, X. Zhang, and D. Zhao, "The CAS-PEAL large-scale Chinese face database and baseline evaluations," *IEEE Trans. on Systems, Man and Cybernetics, Part A: Systems and Humans*, vol. 38, no. 1, pp. 149 –161, 2008.

[52] P. J. Phillips, H. Moon, S. A. Rizvi, and P. J. Rauss, "The FERET evaluation methodology for face-recognition algorithms," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 22, no. 10, pp. 1090–1104, 2000.

[53] Z. Wang, A. Bovik, H. Sheikh, and E. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE Trans. on Image Processing*, vol. 13, no. 4, pp. 600 –612, 2004.

[54] R. Gross, I. Matthews, J. Cohn, T. Kanade, and S. Baker, "Multi-pie," *Image and Vision Computing*, vol. 28, no. 5, pp. 807–813, 2010.

[55] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

Jingang Shi received the B.S. degree and Ph.D. degree in the Department of Information Engineering from the School of Electronic and Information Engineering, Xi'an Jiaotong University, Xi'an, China. He is currently a postdoctoral researcher at the Center for Machine Vision and Signal Analysis, University of Oulu, Finland. His current research topics include image super-resolution, face analysis and biomedical signal processing.

Guoying Zhao is currently a Professor with the Center for Machine Vision and Signal Analysis, University of Oulu, Finland and a professor with the School of Information and Technology, Northwest University, China. She received the Ph.D. degree in computer science from the Chinese Academy of Sciences, Beijing, China, in 2005. She has authored or co-authored more than 190 papers in journals and conferences. Her papers have currently over 9000 citations in Google Scholar (h-index 43). Her current research interests include image and video descriptors, facial-expression and micro-expression recognition, dynamic-texture recognition, human motion analysis, and person identification. Dr. Zhao was a Co-Chair of many International Workshops at ECCV, ICCV, CVPR, ACCV and BMVC. She is co-publicity chair for FG2018, has served as Area Chairs for several conferences. Currently, she is Associate Editor for *Pattern Recognition*, *IEEE Transactions on Circuits and Systems for Video Technology*, and *Image and Vision Computing* Journals.