Received: 24 April 2020 | Revised: 30 March 2021 | Accepted: 29 April 2021

IET Image Processing

DOI: 10.1049/ipr2.12269

The Institution of Engineering and Technology WILEY

ORIGINAL RESEARCH PAPER

Face hallucination based on cluster consistent dictionary learning

Minqi Li[1] | Xiangjian He[2] | Kin-Man Lam[3] | Kaibing Zhang[1] | Junfeng Jing[1]

[1] School of Electronics and Information, Xi'an Polytechnic University, Xi'an 710048, China

[2] Faculty of Engineering and Information Technology, University of Technology, Sydney, NSW 2007, Australia

[3] Centre for Signal Processing, Department of Electronic and Information Engineering, The Hong Kong Polytechnic University, Kowloon, Hung Hom, Hong Kong

Correspondence
Kaibing Zhang, School of Electronics and Information, Xi'an Polytechnic University, Xi'an 710048, China.
Email: zhangkaibing@xpu.edu.cn

Funding information
National Natural Science Foundation of China, Grant/Award Numbers: 61471161, 61971339; Key Project of the Natural Science Foundation of Shaanxi Province, Grant/Award Number: 2018JZ6002; Doctoral Startup Foundation of Xi'an Polytechnic University, Grant/Award Numbers: BS1616, BS1726

Abstract

Face hallucination is a super-resolution technique specially designed to reconstruct high-resolution faces from low-resolution faces. Most state-of-the-art algorithms leverage position-patch prior knowledge of human faces to better super-resolve face images. However, most of them assume the training face dataset is sufficiently large, well cropped or aligned. This paper, proposes a novel example-based face hallucination method, based on cluster consistent dictionary learning with the assumption that human faces have similar facial structures. In this method, the paired face image patches are firstly labelled as face areas including eyes, nose, mouth and other parts, as well as non-face areas without requiring the training face images cropped and aligned. Then, the training patches are clustered according their labels and textures. The cluster consistent dictionary is learned to represent the low-resolution patches and the high-resolution patches. Finally, the high-resolution patches of the input low-resolution face image can be efficiently generated by using the adjusted anchored neighbourhood regression. As utilizing the labelled facial parts prior knowledge, the proposed method represents more details in the reconstruction. Experimental results demonstrate that the authors' algorithm outperforms many state-of-the-art techniques for face hallucination under different datasets.

1 | INTRODUCTION

Face image related techniques have been well developed and investigated in recent years. These techniques have been widely used in many applications such as face recognition, video surveillance, facial expression recognition, digital entertainment, 3D face modelling, and so on. However, due to the limitations of capturing systems and the changes of environment, human face images captured are very often of low resolution. The poor quality of face images has adverse effect on the performances of computer vision and pattern recognition applications. To solve the problem, it is necessary to render a high-resolution (HR) face image from the corresponding low-resolution (LR) one. This technique is named face hallucination (FH) or face super-resolution (SR) [1, 2].

The major difference between face hallucination and the general super-resolution problem is that the face images have regular structures and textures. Compared with the general super-resolution problem, face hallucination is challenging because people are sensitive to the changes in appearances and the quality of human face images. Small deviations might significantly affect human perception, whereas for super-resolution of generic images, such as buildings, plants, etc. the errors can be more tolerant [3]. Another challenge of hallucinating face images is that faces may be in complex conditions, such as under variations of illumination, pose, and expression. Furthermore, it is difficult to align faces in LR images [1, 3].

Different strategies have been researched for face hallucination, such as interpolation-based, degrading model-based, example-based methods and so on. Because of the simplicity of interpolation, it is applied in those applications with low requirements. However, this parametric method is often unable to interpolate details well, such as texture and corner-like local regions [3, 4]. Compared to the interpolation methods, using edge-statistical information can well reconstruct edge and corner areas. For example, Fattal and Raanan [5] imposed edge statistics for image up-sampling. Sun et al. [6] proposed an image super-resolution method by using edge and primal

This is an open access article under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.
© 2021 The Authors. *IET Image Processing* published by John Wiley & Sons Ltd on behalf of The Institution of Engineering and Technology

sketch priors. The major drawback of using edge priors is that they focus on preserving edges so the performance in relatively smooth regions is mediocre [2, 7]. Yang et al. combined a landmark localization method and the gradient map to estimate and align facial features for hallucination [8]. Gradient profile prior was used to enhance the quality of the hallucinated HR image [9]. However, this method is strongly dependent on the results of landmark localization.

Example-based super-resolution schemes have proven to be able to reconstruct significantly finer details from an LR image compared with the interpolation-based schemes [4]. The general idea of example-based approaches is to learn the statistical correlation between pairs of LR and HR images from a face dataset. The learned correlation is then applied to an input LR image to reconstruct the corresponding HR image [3]. Different methods have been studied to learn the mapping relationship between LR and HR images [1, 10, 11], such as

1) Sparse representation-based approaches [12–14];
2) Subspace learning approaches, including local linear embedding and linear subspace learning-based approaches [15–18];
3) Bayesian inference approaches: learning priors from numerous feature vectors to generate a function, mapping features from LR images to HR images [4, 16, 19].

Performance of learning-based SR methods heavily depends on the similarity between the training and the testing images to query input LR face images. The quality of the edges in a reconstructed HR image can be significantly degraded when the edges in training images cannot be matched or aligned well with the corresponding input image [20].

Many researchers have presented that the structural constraints can be applied to improve the results of face hallucination. For example, Markov random fields can be used to reduce the ambiguity between LR and HR images by learning the statistical relationship between a global face image and its local features [3]. Face image structures like facial components are exploited to transfer the high-frequency details for preserving the structural consistency [8].

Similarly, the position information about face images can be used to improve the face hallucination performance. Ma et al. synthesized the high-resolution image patch using the same position image patches of training image pairs [21]. Similar strategy was also proposed in [22] by using convex optimization. Jiang et al. proposed a face super-resolution via locality-constrained neighbour representation based on the position information [23] and contextual information [24]. Lie et al. presented a robust locality-constrained bi-layer representation model to hallucinate the face images [25]. Lu et al. proposed manifold-regularized group locality-constrained representation (MGLR) to exploit the multiple manifold structures rooted in grouped self-similarly patches [26].

From these methods, we can see that faces are highly structured and the positions information of patches from facial images are greatly concerned to improve the face hallucination performance by getting the same face patches' positions. However, most of the existing methods require the faces to be well cropped and accurately aligned, which are challenging tasks, especially for real world face images. In addition, they ignore a fact that only using a single patch with local constraints may result in unstable solutions [26].

Recently, deep learning based super resolution methods have been proposed and claimed the state-of-the-art performance. The pioneer work proposed by [27] is termed as SRCNN, which learns an end-to-end mapping between the bicubic interpolated LR images and the HR images. To get better performance, Dong et al. redesigned the SRCNN structure as FSRCNN by introducing a deconvolution layer at the end of the network [28]. Smaller filter sizes but more mapping layers were adopted to speed up the method. Yamanaka et al. proposed a model with skip connection and network in network (DCSCN) to improve the efficiency [29]. Kim et al. used a very deep convolutional network based on VGG-net to improve the accuracy of super resolution problem [30]. Li et al. proposed a feedback network (SRFBN) to refine low level representations with high-level information [31]. Similarly, structure information was considered in the deep learning framework. For example, Lu et al. developed a parallel region based deep residual network (PRDRN) to predict the missing detailed information for accurate face hallucination [32]. Usually, deep learning based methods demand a large training dataset, intensive computation and memory resources [29].

On the other hand, generative adversarial network (GAN) was proposed for super resolution and face hallucination problems recently. The seminal work proposed by [33] was capable of generating realistic textures during single image super resolution. Wang et al. introduced residual-in-residual dense block (RRDB) to improve the performance of the original SRGAN. The discriminator predicted relative realness instead of the absolute value for better visual quality [34]. Yu et al. proposed transformative discriminative neural networks to avoid heavily relying on accurate alignment of low-resolution (LR) faces before upsampling them [20]. However, the hallucinated face details by GAN based algorithms are often accompanied with unpleasant artifacts.

Inspired by the successful application of example-based learning methods for the super resolution problem and highly structural characters of face images, in this paper, we propose a novel example-based face hallucination method based on cluster consistent dictionary learning. Traditional dictionary learning algorithms [10, 14, 15] focus on the best sparse representation for the training signals of the learned dictionary, but do not consider the consistent capability of the dictionary [36]. Specifically, in the face hallucination problem, similar image signals (in feature spaces) may be represented by different atoms in the dictionary. However, in the test stage, we strongly expect the test image signal has a similar sparse representation to a training sample if they are close in a feature space from the same face part (or in the same cluster). This is particularly useful for face hallucination problems, as face images have similar structures including eyes, nose, mouth etc. Unfortunately, most dictionary learning methods have not considered the cluster consistency. In this project, we train a cluster consistent K-SVD (CC K-SVD) dictionary combined with the adjusted anchored

neighbourhood regression [18] for the face hallucination problem.

The main contributions are summarized as follows: we study the face patches clustering by both the texture similarity and face parts positions. The generated clusters benefit to construct consistent sparse dictionary. We develop a novel example-based face hallucination method, based on discriminative cluster consistent dictionary learning, to exploit facial parts similarity prior without requiring the training face images cropped and aligned. Extensive experimental results on several benchmarks indicate that, by utilizing the prior knowledge of labelled facial parts, our proposed method represents more details in the reconstruction from a small training dataset than many state-of-the-art methods.

The remainder of this paper is organized as follows. Section II reviews the related dictionary learning methods for super resolution and face hallucination problems. In Section 3, we present the details of our proposed method. In Section IV, implementation details are described, and experimental results are presented to show the performance of our method. Finally, in Section V, we draw a conclusion.

2 | RELATED WORK

Sparse coding has been successfully applied to the super-resolution problem. The performance of sparse coding related applications heavily relies on the quality of the over-complete dictionary D. As our proposed method extends the traditional dictionary learning algorithm for face hallucination problem, in this section, we review the related dictionary-based methods for SR.

2.1 | Sparse coding approaches

Sparse coding approaches try to represent the patches by training a codebook of dictionary atoms.

$$\min_{\alpha} \|\mathbf{x} - \mathbf{D}_l\alpha\|_2^2 + \lambda\|\alpha\|_1, \tag{1}$$

where $\mathbf{D}_l$ is the learned dictionary, $\mathbf{x}$ is the low resolution input patch, λ is a weighting factor, and α is the sparse coefficient of dictionary atoms for $\mathbf{x}$.

In order to construct a high resolution image $\mathbf{y}$, the LR and HR dictionaries are jointly trained so that they can represent HR patches and their corresponding LR counterparts using one sparse representation [13, 14, 35]. For example,

$$\min_{D_l, D_h, \alpha} \|\tilde{\mathbf{D}}\alpha - \tilde{\mathbf{Y}}\|_2^2 + \lambda\|\alpha\|_1, \tag{2}$$

where $\tilde{\mathbf{D}} = \begin{bmatrix} \mathbf{D}_l \\ \beta\mathbf{D}_h \end{bmatrix}$ and $\tilde{\mathbf{Y}} = \begin{pmatrix} \mathbf{X} \\ \beta\mathbf{Y} \end{pmatrix}$,$\mathbf{D}_l$, $\mathbf{D}_h$ are the LR, HR dictionary respectively, α is the sparse representation for both $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_N]$ and $\mathbf{Y} = [\mathbf{y}_1, \ldots, \mathbf{y}_N]$ denoting the LR and HR image patches pairs in the training dataset, λ is a weighing factor to balance the importance of the sparsity regularisation, and β controls the tradeoff between matching the LR input and finding an HR counterpart.

Once the dictionaries are trained, given the optimal solution α^* of input testing LR patch $\hat{\mathbf{x}}$, the high-resolution patch can be easily reconstructed as $\hat{\mathbf{y}} = \mathbf{D}_h\alpha^*$.

2.2 | Adjusted anchored neighbourhood regression

Instead of considering the whole dictionary like the sparse encoding approach, anchored neighbourhood regression (ANR) reformulates the patch representation problem as a least square regression regularised by the l2-norm in local neighbourhoods like,

$$\arg\min_{\alpha} \|\mathbf{x} - \mathbf{N}_l\alpha\|_2^2 + \lambda\|\alpha\|_2, \tag{3}$$

where $\mathbf{N}_l$ is the local neighbourhoods of the dictionary atoms. A projection matrix can be precalculated based on the neighbourhood. Finally, an LR input patch can be projected to HR space as,

$$\hat{\mathbf{y}} = \mathbf{N}_h(\mathbf{N}_l^T\mathbf{N}_l + \lambda\mathbf{I})^{-1}\mathbf{N}_l^T\mathbf{x} = \mathbf{P}_j\mathbf{x}, \tag{4}$$

where $\mathbf{N}_h$ is the local neighbourhoods of HR dictionary; $\mathbf{P}_j$ is the stored projection matrix for dictionary atom $\mathbf{d}_{lj}$.

To improve the reconstruction quality, in adjusted anchored neighbourhood regression (A+), the neighbourhood in terms of the dense training samples rather than the sparse dictionary atoms are used in the ridge regression formulation of ANR. The optimization problem then can be presented as,

$$\arg\min_{\delta} \|\mathbf{x} - \mathbf{S}_l\delta\|_2^2 + \lambda\|\delta\|_2, \tag{5}$$

where matrix $\mathbf{S}_l$ contains training samples that lie closest to the dictionary atom to which the input patch y is matched; δ is the weight vector for representing $\mathbf{x}$. Similar to Equation (4), the regressor can be defined by:

$$\hat{\mathbf{y}} = \mathbf{S}_h((\mathbf{S}_l)^T\mathbf{S}_l + \mu\mathbf{I})^{-1}(\mathbf{S}_l)^T\mathbf{x} = \mathbf{F}\mathbf{x}, \tag{6}$$

$$\mathbf{F} = \mathbf{S}_h((\mathbf{S}_l)^T\mathbf{S}_l + \mu\mathbf{I})^{-1}(\mathbf{S}_l)^T, \tag{7}$$

matrix $\mathbf{S}_l$ and $\mathbf{S}_h$ contain training samples that are most correlative to the corresponding $\mathbf{D}_l$ and $\mathbf{D}_h$ dictionary atoms to which the input patch $\mathbf{x}$ is matched; and μ is the regularization parameter.

Instead of $\mathbf{N}_l$ and $\mathbf{N}_h$, $\mathbf{S}_l$ and $\mathbf{S}_h$ use the full training samples learning the regressors on the dictionary as ANR method, which improves the reconstruction performance.

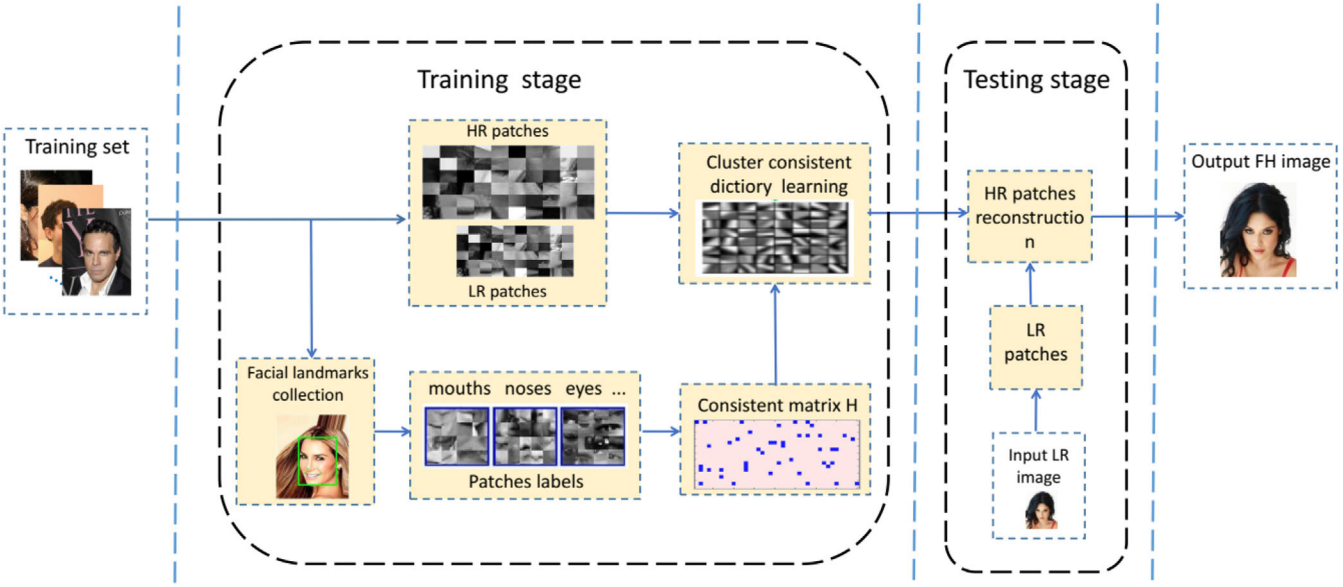


FIGURE 1 The framework of our proposed face hallucination method

3 | PROPOSED METHOD

It is well known that human face images contain complicated local structures. To represent the underlying face geometric structures well, we advocate dividing the processed image patches into several groups such that each group shares similar geometric structures of face attributes. To this end, we utilize positions of facial landmarks to group the positioned training patches. After that, we cluster the patches in each group based on their texture similarities. Then, inspired by the scheme from [36] for discriminative dictionary learning, we design a cluster consistent dictionary learning for the FH problem.

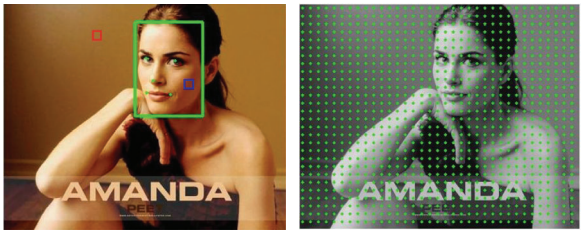
In the traditional dictionary learning based method for SR, the atoms in the dictionary are learned from the training patches independently. These atoms spanned the whole space are used as anchors in the construction step. However, in the testing stage, the similar patches may be expressed by different anchors, especially, when they are located in the boundary of different areas spanned by different anchors. This inconsistent expression of similar feature patches may cause poor results in the reconstruction step for the FH problem. In this paper, we learn a single over-complete dictionary keeping cluster consistent jointly, which yields dictionaries so that face feature patches with the same class labels have similar sparse codes. Similar to advances in the SR family of dictionary learning models and anchored regressors, we also elaborate on designing a set of simple yet efficient linear regressors for FH reconstruction based on the learned CC dictionary to find the underlying local manifold structures such that the learned anchored points can better approximate the subspace of the training dataset.

Figure 1 demonstrates the framework of the proposed face hallucination method. It is comprised of two stages, that is, the training stage and the testing stage. Firstly, we collect training face images with face area label boxes and key feature positions like eyes, noses, mouths etc.. Labelled face datasets [37, 38] or classical face detection methods [39] can be utilized here for getting face area label boxes and key feature positions. Then, a large set of LR and HR paired patches are created from the training set. We divide the patches into five areas as eyes, noses, mouths, other face areas and non-face areas based on the distances between the patch centre positions and the key feature positions. Next, the five areas are clustered respectively. A cluster consistent matrix is constructed by the full clustered patches. Based on the label consistent matrix, different with the traditional K-SVD method, a cluster consistent dictionary is learned to represent the LR and HR patches jointly.

In the testing stage, a given LR image is first divided into the different patches. An LR input patch can be projected to the HR space by the neighbourhood in terms of the dense training samples using ridge regression formulation. Integrating all of the obtained HR patches according to their positions, the final HR image can be generated by averaging pixel values in the overlapping regions.

3.1 | Patches clustering

In order to apply the structure information of face images, we first get the face areas (face boxes) from the training dataset. For those datasets without images with labeled face components, classical face detection algorithms can be used here, such as those in [39–41]. The face components in a face box are grouped as eyes, mouths and noses. For the face areas without face components and background areas, we group them into face-areas or non-face areas, as shown in Figure 2. Specifically, for a patch in a face box area, if the minimum distance between the patch centre and the marked position of a face component is less than 1/4 of the width of the face box, we group the patch



(a) (b)

FIGURE 2 Example of a training image from LFPW face dataset: (a) shows a detected face area described by a green rectangle, and key facial parts are presented in green dots. The blue square patch is labelled as other face area, while the red square patch is labelled as background and (b) shows a uniform sample grid for training patches, where each dot represents the position of the centre of a training patch

to the face component. Otherwise, it is grouped as the normal face area. The left patches which are not inside the face box are grouped into the non-face area or background. Figure 3 shows examples of different groups of patches in facial areas.

Then we cluster the patches separately based on their group labels by traditional clustering algorithms like K-means. We empirically set the number of clusters in each group on different face datasets. More details are described in Section 4. The total cluster number is equal to the summary of the cluster numbers in all groups of different labelled patches.

By performing a clustering algorithm like k-means, we can label the training face patches with K clusters. Each cluster is composed of the patches with similar geometric and texture structures. To well adapt to different contents in an image, once these clusters are formed, we can construct a cluster consistent matrix for the cluster consistent dictionary learning.

3.2 | Cluster consistent dictionary learning

In this section, we describe the construction of a cluster consistent matrix and learning of a cluster consistent dictionary.

We aim to leverage the facial structural information (i.e. position based labels) of input training image patches to learn a reconstructive and label consistent dictionary. The dictionary atoms can reveal different image structures in each cluster, which spans the whole feature space. Each dictionary atom will be chosen so that it represents a subset of the training patches ideally from a single class (cluster), for example, each dictionary item $\mathbf{d}_k$ can be associated with a particular cluster such that representing the corresponding underlying structure.

Consider a collection of N LR and HR image patch pairs in the training dataset, denoted by $\mathbf{X} = [\mathbf{x}_1, \dots, \mathbf{x}_N] \in R^{m \times N}$ and $\mathbf{Y} = [\mathbf{y}_1, \dots, \mathbf{y}_N] \in R^{M \times N}$, m and M are dimensions of LR and HR image patches respectively. To learn a dictionary with K items $\mathbf{D} = [\mathbf{d}_1, \dots, \mathbf{d}_K]$ for sparse representation of $\mathbf{X}$, the consistent matrix H of all the training patches can be defined according to the patch cluster labels.

For example, assuming five training patches $[\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, \mathbf{x}_4, \mathbf{x}_5]$ are from two clusters. Specifically, $\mathbf{x}_1$ and $\mathbf{x}_2$ are from class 1, $\mathbf{x}_3, \mathbf{x}_4$ and $\mathbf{x}_5$ are from class 2. Then the label consistent matrix can be constructed by

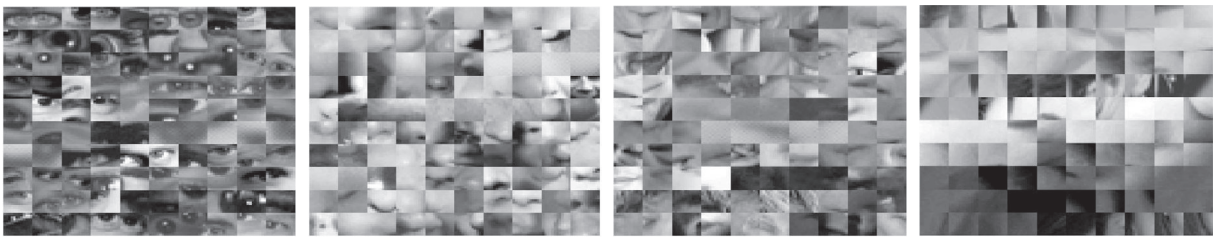
$$H \equiv \begin{bmatrix} 1 & 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 1 \\ 0 & 0 & 1 & 1 & 1 \\ 0 & 0 & 1 & 1 & 1 \end{bmatrix}. \tag{8}$$

We say that $\mathbf{h}_i$ is a cluster label vector of input patches $\mathbf{X}$. The non-zero values of $\mathbf{h}_i$ at those indices indicate that the corresponding patches are from the same cluster.

For obtaining label consistent sparse codes $\mathbf{X}$ with the learned $\mathbf{D}$, an objective function for dictionary construction is defined by:

$$< \mathbf{D}, \mathbf{A}, \alpha > = \underset{\{\mathbf{D},\mathbf{A},\alpha\}}{\arg\min} \|\mathbf{X} - \mathbf{D}\alpha\|_2^2 + \lambda \|\mathbf{H} - \mathbf{A}\alpha\|_2^2, \quad s.t.\ \forall i,\ \|\alpha_i\|_0 \le T. \tag{9}$$

The parameter λ controls the tradeoff between the reconstruction error and the label consistent regularization, T is a sparsity constraint factor. $\mathbf{H} \in R^{M \times N}$ is the cluster label consistent matrix. $\mathbf{A}$ denotes a linear transformation matrix. If the non-zero values $\mathbf{h}_i$ occur at those positions, then the corresponding input patches from $\mathbf{X}$ and the dictionary items from $\mathbf{D}$ share the same label. The first term represents the reconstruction error, while the second term represents the cluster label consistent error, which enforces that the sparse codes α approximate the label consistent matrix sparse codes $\mathbf{H}$ and forces the patches from the same class to have very similar sparse



(a) eyes labelled patches (b) noses labelled patches (c) mouths labelled patches (d) other face parts labelled patches

FIGURE 3 Labelled face patches on LFPW dataset. (a) A group of eyes labelled patches, (b) a group of noses labelled patches (c) a group of mouths labelled patches and (d) group of other face parts labelled patches

representations (i.e., encouraging label consistency in the resulting sparse codes).

Considering that only a few atoms that are closely correlated to the input contribute to the representation, it is reasonable to divide the whole feature space into different groups, such that the atoms in each group are closely correlated to each other. Therefore, the dictionary learned in this way will be adaptive to the underlying face local structure of the training data (leading to a good representation for each member in the set with strict sparsity constraints), and will generate consistent sparse codes α regardless of the size of the dictionary. The anchored points from the dictionary can better approximate the subspace of the training dataset. In the next section, we will show that the consistent property of sparse code α benefits the performance of face reconstruction.

3.3 | Optimization

Traditional sparse coding methods for super resolution applications can use K-SVD algorithm to find the optimal solution for all parameters simultaneously in Equation (2) [14]. K-SVD performs dimensionality reduction on the patches through PCA and using orthogonal matching pursuit (OMP) for the sparse coding. For our face hallucination application, we learn dictionary D, A and coefficient α simultaneously. From Equation (2), Equation (9) can be rewritten to,

$$\begin{aligned} <\hat{\mathbf{D}}, \mathbf{A}, \alpha> &= \arg\min_{\{\hat{\mathbf{D}}, \mathbf{A}, \alpha\}} \|\hat{\mathbf{y}} - \hat{\mathbf{D}}\alpha\|_2^2 + \lambda\|\mathbf{H} - \mathbf{A}\alpha\|_2^2 \\ &= \left\| \begin{pmatrix} \hat{\mathbf{y}} \\ \sqrt{\lambda}\mathbf{H} \end{pmatrix} - \begin{pmatrix} \hat{\mathbf{D}} \\ \sqrt{\lambda}\mathbf{A} \end{pmatrix} \alpha \right\|_2^2, \\ & s.t.\ \forall i,\ \|\alpha_i\|_0 \le T, \end{aligned} \tag{10}$$

where $\hat{\mathbf{D}} = \begin{pmatrix} \mathbf{D}_l \\ \beta\mathbf{D}_h \end{pmatrix}$ and $\hat{\mathbf{y}} = \begin{pmatrix} \mathbf{X} \\ \beta\mathbf{Y} \end{pmatrix}$. The parameter β controls the tradeoff between matching the LR input and finding an HR patch that is compatible with its neighbours. In all of our experiments, we simply set $\beta = 1$.

Let $\widetilde{\mathbf{Y}} = (\hat{\mathbf{y}}; \sqrt{\lambda}\mathbf{H})^T$, $\widetilde{\mathbf{D}} = (\hat{\mathbf{D}}; \sqrt{\lambda}\mathbf{A})^T$, the optimization of Equation (10) is equivalent to solving the following problem:

$$\begin{aligned} <\widetilde{\mathbf{D}}, \alpha> &= \arg\min_{\widetilde{\mathbf{D}}, \alpha} \|\widetilde{\mathbf{Y}} - \widetilde{\mathbf{D}}\alpha\|_2^2, \\ & s.t.\ \forall i,\ \|\alpha_i\|_0 \le T. \end{aligned} \tag{11}$$

This is the classical problem that K-SVD solves [42]. Following K-SVD, we can learn $\widetilde{D}, \alpha$ (e.g., $\hat{D}, A, \alpha$) simultaneously. We use the efficient K-SVD algorithm to find the optimal solution for all parameters, which produces a label consistent sparse representation regardless of the size of the dictionary.

Let $E_k = (Y - \sum_{j \ne k} d_j \alpha^j)$, where α^j is the jth row in α, d_j is the jth column of dictionary $\mathbf{D}$. Let $\tilde{E}_k$, $\tilde{\alpha}_k$ denote the result of discarding the zero items in E_k and α_k, respectively. d_k and $\tilde{\alpha}_k$ can be estimated by solving the following equation [12]:

$$d_k, \tilde{\alpha}_k = \arg\min_{d_k, \tilde{\alpha}_k} \{\tilde{E}_k - d_k\tilde{\alpha}_k\}.$$

Applying SVD decomposition $\tilde{E}_k = \mathbf{U}\Sigma\mathbf{V}^T$, then d_k and $\tilde{\alpha}_k$ are computed as:

$$d_k = \mathbf{U}(:, \mathbf{1}), \tag{12}$$

$$\tilde{\alpha}_k = \Sigma(1,1)V(:,1). \tag{13}$$

3.4 | Training and test

At the training stage, given a set of HR face images, the corresponding LR images are generated by using a bicubic kernel function. We randomly extract a large set of HR and LR patches from the HR and LR image pairs to form a training set.

The patches represented as feature vectors with mean values subtracted are assigned to different clusters based on their textures and facial structures as described in Section 3.1. We expect that the patches in the same cluster have a similar distribution, which does not require the face images to be strictly cropped or aligned. Then, we learn a sparse dictionary $\mathbf{D}_l$ and its corresponding $\mathbf{D}_h$ by enforcing the same coefficients and cluster consistency in the HR and LR patch decompositions over $\mathbf{D}_l$ and $\mathbf{D}_h$ as Equation (11).

Instead of considering the whole dictionary like the sparse encoding approach, local neighbourhoods of the dictionary or neighbourhoods of training samples have been proven to provide better reconstruction quality [18]. Therefore, we follow the A+ strategy to reconstruct HR patches by reusing the training samples. Specifically, for each dictionary atom $\mathbf{d}_k^l, k \in (1, K)$, we define the neighbourhood $\{\mathbf{S}_k^l, \mathbf{S}_k^h\}$ in terms of the dense training samples rather than the sparse dictionary atoms. Then, the regressor can be defined by Equation (6).

For testing, the input LR image is divided into patches. Each LR patch is represented by the learned cluster label consistent dictionary. When all of the K regressors corresponding to the K clusters of patches obtained, for a given testing LR patch $\mathbf{x}_l$, the most correlated dictionary atom $\mathbf{d}_k^l$ and its corresponding regressor are applied to estimate the output $\mathbf{y}$ by

$$\hat{\mathbf{y}} = \mathbf{F}_j\mathbf{x}_l, \tag{14}$$

where $\mathbf{F}_j$ is the most matched regressor with $\mathbf{x}_l$, which is measured by the maximal absolute value of correlation between $\mathbf{x}_l$ and the atoms cross the dictionary $\mathbf{D}_l$, that is

$$j = \underset{j=1,\ldots K}{\operatorname{argmax}} \quad [abs((d_j^l)^T\mathbf{x}_l)], \tag{15}$$

where $abs(\cdot)$ denotes the absolute function.

Finally, the HR face image is reconstructed by averaging the overlapped areas of the HR patches. The whole process is shown in Algorithm 1.

ALGORITHM 1 Face hallucination by label consistent dictionary learning

Input: Initial LR training set and corresponding HR training set; LR test image $\mathbf{x}_l$.

Training stage:

1: Divide the LR and HR training faces into overlapping patches $\mathbf{X} = [\mathbf{x}_1, \dots, \mathbf{x}_N]$ and $\mathbf{Y} = [\mathbf{y}_1, \dots, \mathbf{y}_N]$ with the corresponding patches centre positions $\mathbf{P}_l = [\mathbf{p}_l^1, \dots, \mathbf{p}_l^N]$ and $\mathbf{P}_h = [\mathbf{p}_h^1, \dots, \mathbf{p}_h^N]$.
2: Group the patches based on the distances between the patch center and the face land marks under predefined threshold θ.
3: Get clusters labels in each group of patches by K-means clustering.
4: Construct cluster consistent matrix $\mathbf{H}$ via Equation (8).
5: Learn $\mathbf{D}_l$, $\mathbf{D}_h$ and sparse coefficients α via Equations (12) and (13).
6: Build the regressors $\mathbf{F} = [\mathbf{F_1}, \dots, \mathbf{F_K}]$ via Equation (7).

Testing stage:

1: Patition the input test image $\mathbf{x}_l$ into overlapping patches $\mathbf{x}_l^1, \dots, \mathbf{x}_l^n$.
2: Find the best matched regressor F_j for each patch, and compute the HR patch via Equations (15) and (14).

Output: HR face image $\hat{\mathbf{y}}$ by integrating all the HR patches according the positions and averaging overlapping regions.

The key difference between our method and A+ method is: our dictionary gives patches in the same cluster a similar sparse coefficient, for example, a consistent reconstructed result. As only a few atoms that are closely correlated to the input feature vector to the sparse representation, when we choose suitable correlative neighbours of each LR atom from a learned LR and HR dictionary pair, the atoms in the learned CC dictionary can reveal different face image structures. We find such CC dictionary can better represent face details in the face hallucination application. The main reason is the patches from the same face parts have a similar sparse representation in our proposed method. Therefore, they can be consistently reconstructed by the anchored training patches, which give us better face hallucination performance. This is particularly useful for face key parts areas, as eyes, nose and mouth etc.

4 | EXPERIMENTS

To illustrate the performance of the proposed method, we evaluate our algorithm on three popular face databases: FEI face database [43], CelebA database [37] and LFPW database [38].

We compare our method with some classical and state-of-the-art face hallucination and related image super resolution methods including bicubic, sparse coding Yang's [12], A+ [18], FSRCNN [28], TRNR [44], SRGAN [33], TLcR-RL [24], ESRGAN [34], SRFBN [31], and CCR [47]. For all the compared methods, we have retrained the models using the same training dataset as our proposed method. PSNR (valuated on the luminance channel in YCbCr color space for color images) and SSIM are used as the objective measurements of the image quality.

4.1 | Implementation and parameters setting

In this section, we describe the implementation details and the main parameters of our proposed method. Since A+ is the closest related method of ours, as a patch-based method, the training image patch is recommended with size of 12×12, and the overlap between two adjacent patches is suggested to be 1 pixel. We set the dictionary with 1024 anchored points and 2048, the correlative neighbourhood size, for example, $p = 2048$ for fast training as suggested by [18, 45]. We set $\lambda = 0.001$ empirically for controlling the tradeoff between the sparse dictionary learning and the label consistence regularization. As some deep learning based methods only provide a training code on scale factor 4, without loss of generality, we only compare the results on upscale factor 4.

Before training, we first detect the face parts by classical detection methods. Without loss of generality, we extract the key part positions including eyes, noses, mouths, and face boxes by MTCNN [39], which is an efficient and effective open-source tool for face detection. For those databases like CelebA and LFPW, the images have been well labelled. The key part positions can also be extracted based on landmark points from the ground truth.

Then, the patches are grouped into face areas or background (non-face) areas based on their positions to the face boxes. Different from the methods for preparing traditional dictionary training patches like KSVD [46], in this paper, the patches are combined with position information for further labelling. In the labelling step, the distances between the position of each patch and the positions of key parts in each image are computed. We compare the minimum distance with a predefined threshold (e.g., the average distance between left and right eyes). If the minimum distance is less than the predefined threshold θ, the patch is labelled as the corresponding face part. Specifically, if the minimum distance between the centre of the patch and the key part is less than $1/4$ width of the face box, the patch is labelled corresponding to the nearest face key part group. Otherwise, the unlabelled patches are grouped into the face area or the background (non-face) area according the position in or out the predicted face box.

Depending on the labels, we cluster the patches in each group separately based on their textures. The number of cluster in each group is set empirically. More details about the effect from the cluster number K and the consistency controlling parameter λ will be discussed later.

4.2 | FEI database

The FEI face database is composed of 400 frontal face images of 200 persons, for example, two images per person with smiling expression and neutral expression respectively.

In our experiments, the training set contains 360 face images, which are randomly selected from the FEI database. For fair comparison, we do not use data augment technology as applied in GAN based methods like [31]. 1,300,000 LR and HR face patch pairs are collected for training. The remaining images are used for testing. Without loss of generality, we magnify the input face image with a factor of 4. In other words, the original images form the HR image dataset, which are down-sampled with a factor of 4 to form the LR image dataset. We set four clusters

TABLE 1 PSNR/SSIM performances of different algorithms on FEI database

Images	Bicubic	Yang's	A+	FSRCNN	TRNR	SRGAN	SRFBN	ESRGAN	TLcR-RL	CCR	Proposed
PSNR	30.91	32.87	33.7	31.62	33.12	33.19	33.21	31.5635	34.19	33.88	**34.27**
SSIM	0.8715	0.9038	0.9189	0.9204	0.9222	0.8797	0.8805	0.8951	**0.9370**	0.9232	0.9247

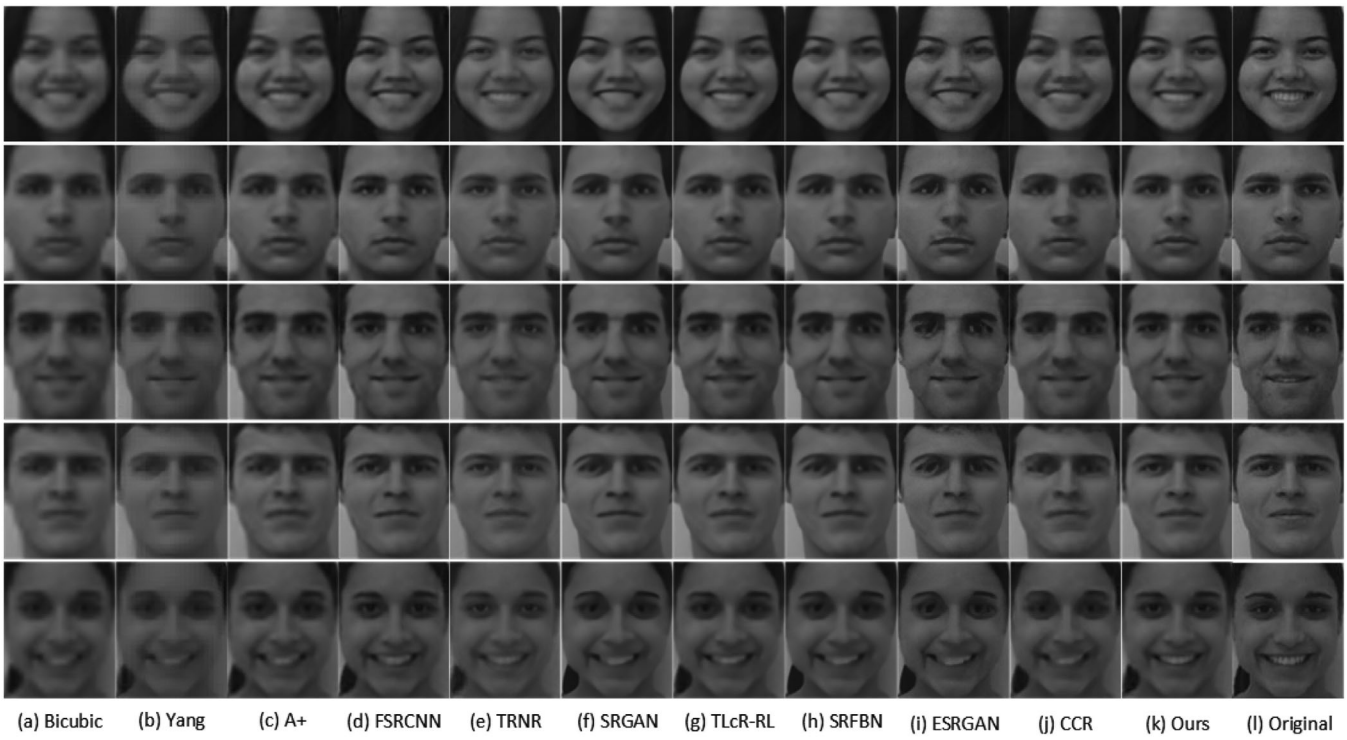


FIGURE 4 Face hallucination results on the FEI dataset for scale 4 with different methods: (a) Bicubic interpolation, (b) Yang's method , (c) A+, (d) FSRCNN, (e) TRNR, (f) SRGAN, (g) TLcR-RL, (h) SRFBN, (i) ESRGAN, (j) CCR, (k) our method and (l) the original HR faces

for eyes, noses and mouths patches, 20 for other face areas, and eight for background.

The quantitative comparison based on different methods is shown in Table 1. The visual results are shown in Figure 4. From Table 1, we can see that the simple bicubic interpolation method cannot produce more high frequency details. Our algorithm performs better than classical dictionary based methods such as Yang's, A+, as well as clustering and collaborative representation method with a margin of improvement of 0.36 in PSNR and 0.01 in SSIM.

Both TNRN and TLcR-RL apply a simple neighbour representation for face hallucination problem. Concretely, TNRN uses a simple neighbour representation with Tikhonov regularization and position information. TLcR-RL uses context information and reproducing learning by adding the hallucinated HR face image to the training set. TLcR-RL presents competitive results on the FEI dataset, however, they require the training images well cropped (not suitable for FLPW dataset as described in in the following section) and the reproducing learning in the reconstruction step is time consuming.

As for the deep learning based methods, such as FSRCNN and SRFBN, we retrain them with FEI. Unfortunately, FSRCNN shows blurry results on facial images with a upscaling factor 4. SRFBN can well maintain the face contours due to their global optimization scheme. However, it fails to capture high frequency details (refer to the eyes, noses and mouths).

When compared with GAN based methods, SRGAN and ESRGAN can be seen as the currently most popular super resolution and face hallucination methods. Although some results from GAN based methods present better perceptual loss for super resolution of natural images, particularly, ESRGAN achieves relatively sharper face contours, they tend to bring artifacts for human faces. For example, human faces, such as the eyes parts are shown in columns (f) and (i) of Figure 4.

4.3 | CelebA database

To further examine our algorithm, we also conduct the same experiment on a large-scale real face CelebA dataset [37].

TABLE 2 PSNR/SSIM performances of different algorithms on CelebA database

Images	Bicubic	Yang's	A+	FSRCNN	TRNR	SRGAN	SRFBN	ESRGAN	TLcR-RL	CCR	Proposed
PSNR	28.33	29.54	29.94	28.76	28.91	30.06	29.81	28.51	29.58	29.95	**30.17**
SSIM	0.8132	0.8411	0.8512	0.8498	0.8159	**0.8687**	0.8483	0.8247	0.8445	0.8530	0.8617

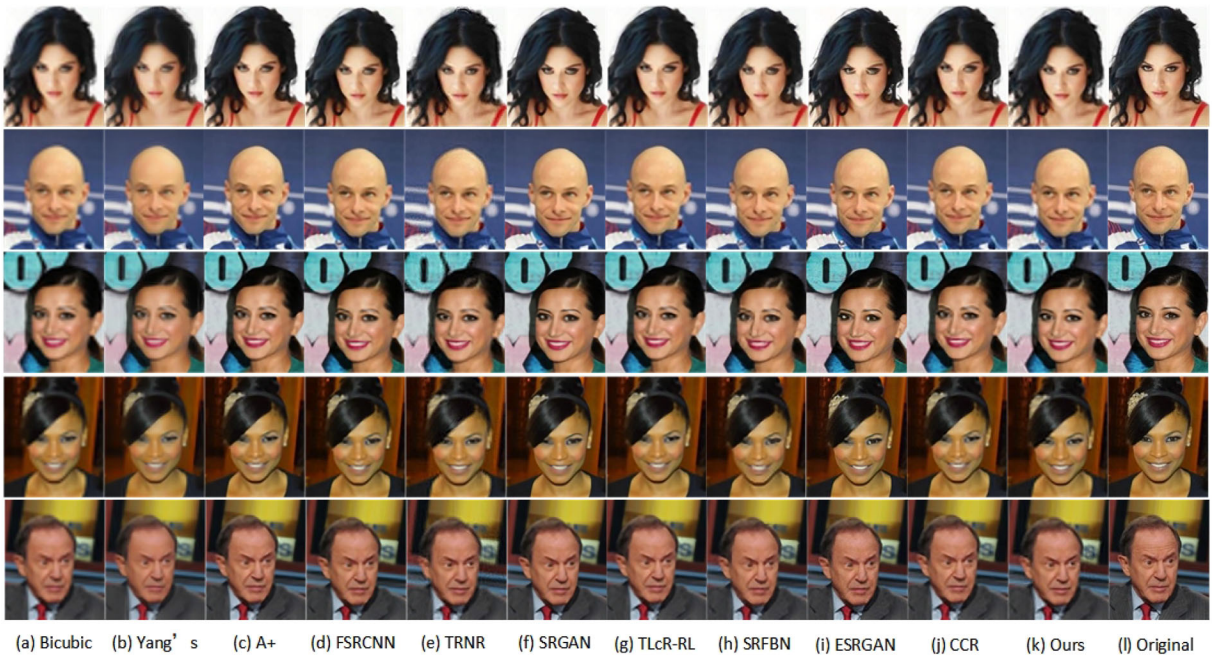


FIGURE 5 Face hallucination results on the CelebA dataset for scale 4 with different methods: (a) Bicubic interpolation, (b) Yang's method , (c) A+, (d) FSRCNN, (e) TRNR, (f) SRGAN, (g) TLcR-RL, (h) SRFBN, (i) ESRGAN, (j) CCR, (k) our method and (l) the original HR faces

CelebA dataset consists of 20,000's of face images, and each image is labelled with five landmarks (two eyes, nose and mouth corners). We randomly select 250 images as training set. For the testing data, 100 images are randomly chosen in the remaining images.

As shown in Table 2 and Figure 5, traditional interpolation, dictionary learning and example based upsampling methods, that is, Yang's, A+, TRNR and TLcR-RL cannot hallucinate clear facial details. Particularly, the sparse coding based super-resolution methods Yang's and A+ may reconstruct similar image signals (in feature spaces) by different atoms in the dictionary without considering finding a consistent correspondence between LR and HR patches. CCR uses local geometry property by clustering to improve the performance. However, the improvement is small. For TRNR and TLcR-RL methods, we adapt the original public codes for colour images. As the samples in CelebA dataset including more different poses are not well aligned as FEI dataset, face structure depended TRNR and TLcR-RL methods cannot present a good performance as shown on FEI dataset. With the help of the discriminator network, SRGAN and ESRGAN methods can achieve good perceptual loss on CelebA dataset. However, the PSNR and SSIM performances are mediocre and they may hallucinate distorted facial details especially in eyes areas. For deep convolutional network based methods SRFBN shows better performance than FSRCNN. However, the performances on the structural CelebA face images by SRFBN are not as good as shown on natural images. Besides, both of them with a large number of parameters, require more qualified hardware support (e.g. GPU) and training samples. This also implies that our up-sampling method is more suitable for the face hallucination task.

4.4 | LFPW database

To further examine the robustness of our algorithm, we also conduct the same experiments on the labelled face parts in-the-wild (LFPW) face database without cropping and alignment [38]. The LFPW database contains 1,432 images with different sizes downloaded from the websites such as google.com, yahoo.com, and flickr.com with large variations in pose, expression, illumination and occlusion. Each image is labelled with 35 landmark points. We randomly select and downsample 100 images as training set. For the testing data, 100 images are randomly chosen in the remaining images.

TABLE 3 PSNR/SSIM performances of different algorithms on LFPW database

Images	Bicubic	Yang's	A+	FSRCNN	TRNR	SRGAN	SRFBN	ESRGAN	TLcR-RL	CCR	Proposed
PSNR	28.05	29.21	29.51	28.24	–	29.04	29.38	29.07	–	29.57	**29.59**
SSIM	0.8145	0.8450	0.8510	0.8374	–	0.8263	0.8404	0.8303	–	0.8512	**0.8520**

FIGURE 6 Face hallucination results on the LFPW dataset for scale 4 with different methods: (a) Bicubic interpolation, (b) Yang's method , (c) A+, (d) FSRCNN, (e) SRGAN, (f) SRFBN, (g) ESRGAN, (h) CCR, (i) our method and (j) the original HR faces

As TRNR and TLcR-RL methods require the training samples to be well cropped, and the LFPW database cannot meet this requirement, we only provide the performances of other methods. Comparing with other algorithms, as shown in Table 3 and Figure 6, our method yields better performance in terms of both PSNR and SSIM.

4.5 | Choice of sensitive parameters

An important parameter is the regularization parameter λ in the cluster consistent dictionary learning model in Equation (10). In this experiment, we investigate how the empirical parameter λ affects the performance of the CCFH method. To obtain an optimal parameter to adapt the reconstruction error and the regularization term, we analyse the PSNR and SSIM performances by varying λ in the range from 10^{-6} to 1.

Figure 7 shows the average performances of the PSNR and SSIM scores corresponding to different regularization values on CelebA dataset. Based on the results, we find that the value of λ does affect the cluster consistent based FH. Too small value of λ is insufficient to keep the consistent relationship, while too large value may impact the reconstruction loss results. We can see that the best regularization value corresponding to the highest PSNR and SSIM is around 10^{-3}. Based on the statistical results, we empirically set λ to 10^{-3} in our experiment.

Another important parameter is the number of clusters. In order to choose a reasonable cluster number for better reconstruction, we make an empirical study on how the parameter affects the construction quality by varying the cluster number K within the range from 5 to 75.

The test is conducted on the training dataset CelebA containing the dictionary with 1024 anchored points for the magnification factor as 4. Figure 8 displays the averaged PSNR and SSIM

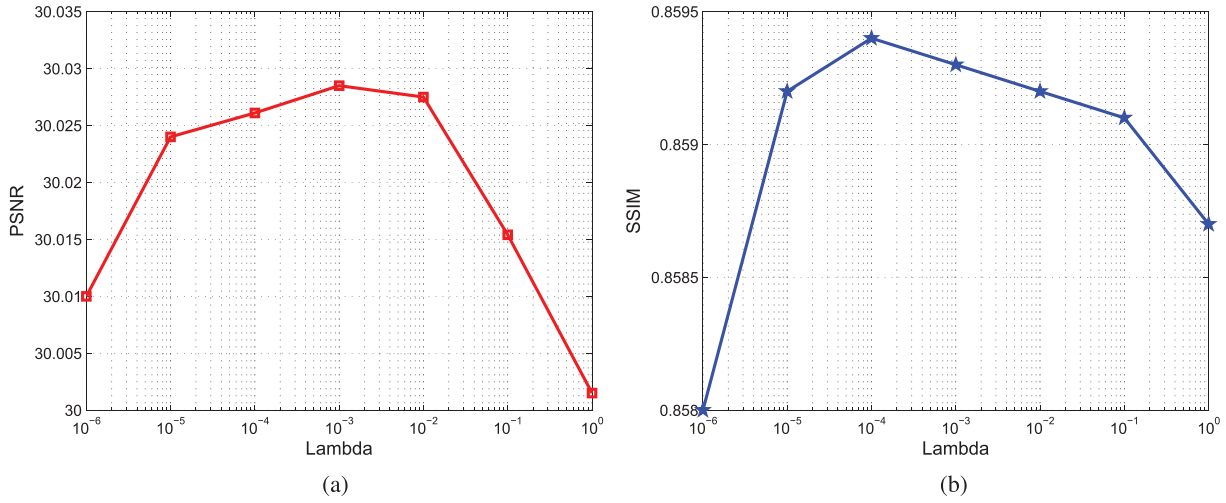


FIGURE 7 Choice of the regularization parameter λ. (a) PSNR results by different values of λ; (b) SSIM results by different values of λ

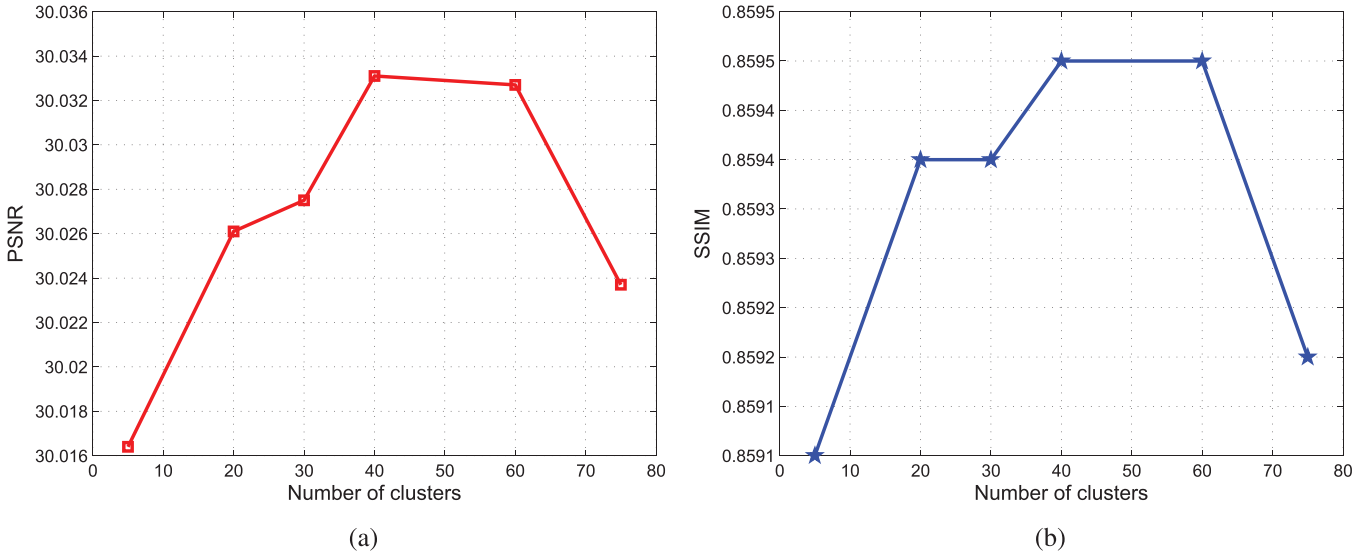


FIGURE 8 The average of PSNR and SSIM scores of CelebA dataset varying with different cluster numbers. (a) PSNR results by different values of the cluster number K; (b) SSIM results by different values of the cluster number K

when different cluster numbers are applied to train the cluster consistent dictionary. Based on the figure, we find that larger number of cluster benefits lower reconstruction error. However, when K is greater than 40, there is no obvious decrease in the reconstruction error. Instead, too many clusters may impact the reconstruction results. According to the experiment, we suggest prefixing $K = 40$ throughout our experiment.

4.6 | Computational complexity analysis

In this section, we discuss the running time of the proposed FH algorithm. Because the running time in the testing stage is the main factor for learning-based FH approaches, here we only focus on the discussion about the computational complexity in the testing stage. Our implementation of CCFH has a similar computation time to that of A+ because of similar strategy, whose time complexity for encoding LR input patches to HR output patches is linear in the number of input image patches and the number of anchoring atoms. Thus, the major procedures of our method involve two parts, that is, the transformation from LR features to HR features and the adaption to estimate the best regressor. The feature transformation from LR to HR using the precomputed mapping matrix costs $O(NmM)$. The computation of finding the most correlative neighbours of N inputs takes $O(NmKp)$ operations by projecting onto the LR dictionary D_l and choosing the most correlative p neighbours with the nearest neighbour searching algorithm. Thus, for our proposed CCFH framework, the total complexity is about $O(Nd_lKp + Nd_ld_h)$.

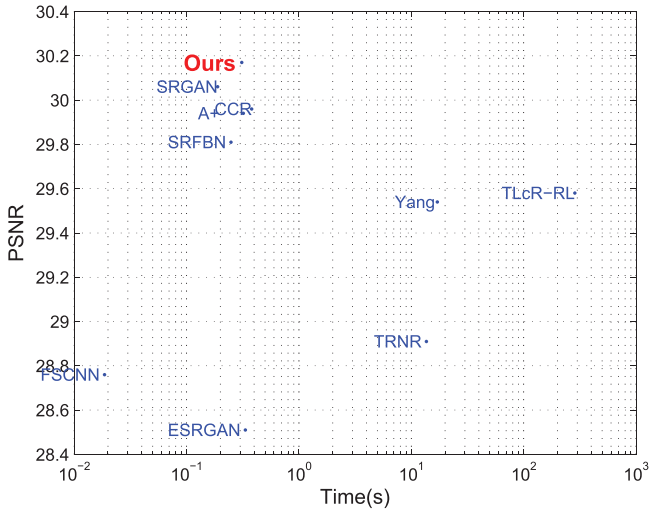


FIGURE 9 PSNR performance versus runtime (evaluated in seconds). The results are evaluated on the CelebA dataset for an upscaling factor of ×4. The proposed method achieves the best performance with relatively less execution time

We further compare the computational efficiency of the different methods. All the comparative experiments are performed on a 1.8 GHz Intel Core i7 CPU with 8 GB RAM and GTX1080 GPU for deep learning based methods. Figure 9 shows the averaged PSNR performance versus runtime (in seconds) evaluated on the CelebA dataset for an upscaling factor of 4. As demonstrated, our method can achieve better FH quality with competitive computational time.

5 | DISCUSSION AND CONCLUSION

In this paper, we have presented a novel example-based face hallucination method based on cluster consistent dictionary learning with the assumption that all face images have similar local pixel structures and similar face image patches should be reconstructed consistently on the same training dictionary. We have grouped the face patches based on their positions to the face key parts and face boxes. Then, all the patches are clustered in each group separately. Cluster consistence matrix can be constructed for learning a constrained dictionary. We have found that such dictionary can better represent face details in the face hallucination application. The main reason is that the patches from same face parts have a similar sparse representation in our proposed method. Therefore, they can be consistently reconstructed by the anchored training patches, giving us better face hallucination performance. This is particularly useful face key parts areas, as eyes, nose, mouth etc.

Experimental results show that the proposed method performs well in terms of both reconstruction error and visual quality. The PSNR and SSIM results of the experiments show that our method can achieve competitive performance for face hallucination.

Moreover, our label consistent method, which is a more flexible constraint to describe the neighbourhood of face image pixels, does not require the training face images well cropped and aligned which is significant for some traditional methods.

ACKNOWLEDGMENT

This work was partially supported by the Chinese National Natural Science Foundation of China under Grant No. 61971339 and Grant No. 61471161, and science and technology planning project of Xi'an under Grant No. 2020KJRC0028, and technology planning project of Beilin, Xi'an under Grant No. GX2006, and in part by the Doctoral Startup Foundation of Xian Polytechnic University under Grant BS1726.

ORCID

Kaibing Zhang https://orcid.org/0000-0002-3770-017X

REFERENCES

1. Wang, N., et al.: A comprehensive survey to face hallucination. Int. J. Comput. Vision 106(1), 9–30 (2014)
2. Li, X. et al.: A face hallucination algorithm via kpls-eigentransformation model. In: IEEE International Conference on Signal Processing, Communication and Computing, Hong Kong, China, pp. 462–467 (2012)
3. Liu, C., Shum, H.Y., Freeman, W.T.: Face hallucination: Theory and practice. Int. J. Comput. Vision 75(1), 115–134 (2007)
4. Hsu, C.C. et al.: Face hallucination using bayesian global estimation and local basis selection. In: IEEE International Workshop on Multimedia Signal Processing, Saint-Malo, France, pp. 449–453 (2010)
5. Fattal, R.: Image upsampling via imposed edge statistics. ACM SIGGRAPH 26(3), 95 (2007)
6. Sun, J. et al.: Image hallucination with primal sketch priors. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Madison, WI, pp. II–729 (2003)
7. Donaldson, K., Myers, G. K.: Bayesian super-resolution of text in videowith a text-specific bimodal prior. Int. J. Doc. Anal. Recogn. 7(2–3), 159–167 (2005)
8. Yang, C.Y., Liu, S., Yang, M.H.: Structured face hallucination. In: IEEE Conference on Computer Vision and Pattern Recognition, Portland, OR, pp. 1099–1106 (2013)
9. Sun, J., Xu, Z., Shum, H.Y.: Gradient profile prior and its applications in image super-resolution and enhancement. IEEE Trans. Image Process. 20(6), 1529–1542 (2010)
10. Timofte, R., Rothe, R., Van Gool, L.: Seven ways to improve example-based single image super resolution. In: IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, pp. 1865–1873 (2016)
11. Chang, H., Yeung, D.Y., Xiong, Y.: Super-resolution through neighbor embedding. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Washington, DC, pp. I–I (2004)
12. Yang, J. et al.: Face hallucination via sparse coding. In: IEEE International Conference on Image Processing, pp. 1264–1267 (2008)
13. Yang, J. et al.: Image super-resolution as sparse representation of raw image patches. In: IEEE Conference on Computer Vision and Pattern Recognition, Washington, DC, pp. 1–8 (2008)
14. Zeyde, R., Elad, M., Protter, M.: On single image scale-up using sparse-representations. In: International Conference on Curves and Surfaces, Avignon, France, pp. 711–730 (2010)
15. Zhang, K., et al.: Partially supervised neighbor embedding for example-based image super-resolution. IEEE J. Sel. Top. Signal Process. 5(2), 230–239 (2010)
16. Li, M. et al.: Face hallucination based on nonparametric bayesian learning. In: IEEE International Conference on Image Processing, Quebec City, Canada, pp. 986–990 (2015)
17. Timofte, R., De Smet, V., Van Gool, L.: Anchored neighborhood regression for fast example-based super-resolution. In: IEEE Conference on Computer Vision and Pattern Recognition, Sydney, Australia, pp. 1920–1927 (2013)

18. Timofte, R., De Smet, V., Van Gool, L.: A+: Adjusted anchored neighborhood regression for fast super-resolution. In: Asian Conference on Computer Vision, Zurich, Switzerland, pp. 111–126 (2014)
19. Pickup, L.C., Capel, D.P., Roberts, S.J., Zisserman, A.: Bayesian image super-resolution, continued. Adv. Neural Inf. Process. Syst. 1089–1096 (2006)
20. Yu, X., Porikli, F.: Face hallucination with tiny unaligned images by transformative discriminative neural networks. In: Proceedings of Thirty-First AAAI Conference on Artificial Intelligence, San Francisco, CA, pp. 4327–4333 (2017)
21. Ma, X., Zhang, J., Qi, C.: Position-based face hallucination method. In: IEEE International Conference on Multimedia and Expo, New York, NY, pp. 290–293 (2009)
22. Jung, C., et al.: Position-patch based face hallucination using convex optimization. IEEE Signal Process. Lett. 18(6), 367–370 (2011)
23. Jiang, J., et al.: Face super-resolution via multi-layer locality-constrained iterative neighbor embedding and intermediate dictionary learning. IEEE Trans. Image Process. 23(10), 4220–4231 (2014)
24. Jiang, J., et al.: Context-patch face hallucination based on thresholding locality-constrained representation and reproducing learning. IEEE Trans. Cybern. 50(1), 324–337 (2018)
25. Liu, L., et al.: Robust face hallucination via locality-constrained bi-layer representation. IEEE Trans. Cybern. 48(4), 1189–1201 (2017)
26. Lu, T. et al.: Face hallucination using manifold-regularized group locality-constrained representation. In: Proceedings 25th IEEE International Conference on Image Processing, Athens, Greece, pp. 2511–2515 (2018)
27. Dong, C. et al.: Learning a deep convolutional network for image super-resolution. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) European Conf. on Computer Vision, Lecture Notes in Computer Science, vol. 8692. Springer, Cham, pp. 184–199 (2014)
28. Dong, C., Loy, C.C., Tang, X.: Accelerating the super-resolution convolutional neural network. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds) European Conference on Computer Vision, Lecture Notes in Computer Science, Springer, Cham, pp. 391–407 (2016)
29. Yamanaka, J., Kuwashima, S., Kurita, T.: Fast and accurate image super resolution by deep cnn with skip connection and network in network. In: International Conference on Neural Information Processing, Bangkok, Thailand, pp. 217–225 (2017)
30. Kim, J. et al.: Accurate image super-resolution using very deep convolutional networks. In: IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, pp. 1646–1654 (2016)
31. Li, Z. et al.: Feedback network for image super-resolution. In: IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, CA, pp. 3867–3876 (2019)
32. Lu, T., et al.: Parallel region-based deep residual networks for face hallucination. IEEE Access 7, 81266–81278 (2019)
33. Ledig, C. et al.: Photo-realistic single image super-resolution using a generative adversarial network. In: IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, pp. 4681–4690 (2017)
34. Wang, X. et al.: Esrgan: Enhanced super-resolution generative adversarial networks. In: European Conference on Computer Vision, Barcelona, Spain, pp. 3637–3641 (2018)
35. Yang, J., et al.: Image super-resolution via sparse representation. IEEE Trans. Image Process. 19(11), 2861–2873 (2010)
36. Jiang, Z., Lin, Z., Davis, L.S.: Learning a discriminative dictionary for sparse coding via label consistent K-SVD. In: IEEE Conference on Computer Vision and Pattern Recognition, Colorado Springs, CO, pp. 1697-1704 (2011)
37. Liu, Z. et al.: Deep learning face attributes in the wild. In: IEEE International Conference on Computer Vision, Santiago, Chile, pp. 3730–3738 (2015)
38. Belhumeur, P.N., et al.: Localizing parts of faces using a consensus of exemplars. IEEE Trans. Pattern Anal. Mach. Intell. 35(12), 2930–2940 (2013)
39. Zhang, K. et al.: Joint face detection and alignment using multitask cascaded convolutional networks. IEEE Signal Process Lett. 23(10), 1499–1503 (2016)
40. Kazemi, V., Sullivan, J.: One millisecond face alignment with an ensemble of regression trees. In: IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, pp. 1867–1874 (2014)
41. Chen, D. et al.: Joint cascade face detection and alignment. In: Fleet D., Pajdla T., Schiele B., Tuytelaars T. (eds.) Computer Vision - ECCV 2014, Lecture Notes in Computer Science. vol 8694, pp. 109–122. Springer, Cham (2014)
42. Elad, M.: Sparse and Redundant Representations: From Theory to Applications in Signal and Image Processing. Springer Science & Business Media, Heidelberg, Germany (2010)
43. Thomaz, C.E., Giraldi, G.A.: A new ranking method for principal components analysis and its application to face image analysis. Image Vision Comput. 28(6), 902–913 (2010)
44. Jiang, J., et al.: Noise robust position-patch based face super-resolution via Tikhonov regularized neighbor representation. Inf. Sci. 367, 354–372 (2016)
45. Zhang, K., et al.: Learning recurrent residual regressors for single image super-resolution. Signal Process. 154, 324–337 (2019)
46. Aharon, M., Elad, M., Bruckstein, A.: K-SVD: An algorithm for designing overcomplete dictionaries for sparse representation. IEEE Trans. Signal Process. 54(11), 4311–4322 (2006)
47. Zhang, Y., et al.: CCR: Clustering and collaborative representation for fast single image super-resolution. IEEE Trans. Multimedia 18(3), 405–417 (2016)

How to cite this article: Li M, He X, Lam Kin-Man, Zhang K, Jing J. Face hallucination based on cluster consistent dictionary learning. IET Image Process.. 2021;15:2841–2853.
https://doi.org/10.1049/ipr2.12269