Face Super-resolution Based on Multi-source References

Rui Wang
School of Computer Science and Technology
Shandong University of Finance and Economics
Jinan, China
capricorn.orz@gmail.com

Muwei Jian*
School of Computer Science and Technology
Shandong University of Finance and Economics
Jinan, China
*Corresponding author:
jianmuweihk@163.com

Paul Smith
School of Criminology and Criminal Justice
University of Portsmouth
Portsmouth, UK
Paul1.Smith@port.ac.uk

Hui Yu
School of Creative Technologies
University of Portsmouth
Portsmouth, UK
hui.yu@port.ac.uk

***Abstract*—This paper proposes a multi-source references (MSR) based face super-resolution (FSR) model. More specifically, to enhance the low-quality large-scale reconstruction of faces without the involvement of face prior knowledge, we propose a multi-source references based FSR framework exploiting a constructed reference library of non-identity faces and an information mining module for external and internal references. Experimental results show that the proposed model can provide more satisfactory and reliable face super-resolution results than the-state-of-the-art methods.**

Keywords—multi-source, face super-resolution, prior knowledge

I. INTRODUCTION

Under controlled conditions (such as outdoor video surveillance), the captured facial images usually have a very low resolution (LR) with different illumination conditions and arbitrary poses. In reality, there are many uncertainties causing different degrees of image quality degradation, which leads to severe deterioration of the performance in a wide range of practical applications. Therefore, there is a need to design a face super-resolution (also called face hallucination) model with greater utility and a larger up-sampling factor.

Compared with the general Single Image Super Resolution (SISR) approach in the field of natural images, face hallucination not only explores the elaborate deep convolutional network architecture, but also considers the inclusion of prior knowledge of faces as a guide during the reconstruction process [1][2], including facial landmarks [3], facial parsing maps and facial heatmaps, etc. Currently, it still has many technical and challenging issues to be solved, particularly when the faces have a very low resolution and arbitrary poses [24, 25, 26].

A lot of reference-based methods have been proposed to achieve robust face hallucination, and those works usually utilize high-resolution reference images (R) of the same identity in multiple scenes as an external guide to capture the available features to enhance the input low-frequency signal. However, they usually have certain requirements for the pose and expression of R, such as a frontal view face with eyes open. In addition, most existing methods are mainly concentrated with a magnification factor of 2 and 4 during face hallucination, which makes it easier for reconstruction but greatly limits the generality of the model.

To solve this problem, a multi-source references based large-scale super-resolution network, called MSRNet, is proposed for facial images. Our contribution can be summarized into the following two points:

1) We propose a reference-based MSRNet for efficient face hallucination with a direct magnification factor of 8;

2) Across-scale adaptive feature matching and fusion modules based on multi-source references are designed to largely address the reliance on a priori knowledge of faces.

In this section, we simply review and discuss recent related work on face image super-resolution, including these methods of designing efficient network structures for LR face images, as well as those models that consider face-specific prior information.

Huang et al. [4] designed a three-layer convolutional network to fit nonlinear mappings for face reconstruction. Later, a Dual-path deep fusion network [5] was proposed for fusing the global information of the face and the details of the local components learned by the two branches, respectively. In [6], Jiang et al. proposed to use the priori nature of facial structure to learn differential evolution-based parameter maps to guide the modeling of co-occurrence relationships in different regions of the input low-resolution (LR) face images. Considering the input LR size is small that accurate prior labels cannot be extracted, FSRNet [7] and MSFSR [8] put this behavior behind coarse super-resolution, and namely the result of coarse LR enhancement is extracted with prior information, and then cascaded with the encoder-extracted features into the next step of fine super-resolution network to generate the final result. Liu et al. [9] proposed a graph representation framework based on modal regression with the aim of avoiding the encoding error of the least square metric and thus enhancing robustness to the noise of uncertain distributions. In [10], Chen et al. proposed homogenized projections in LR space and HR space as a compensation to formulate FSR in a multi-stage framework. Aiming to learn the advantages of locality in preserving the true type of structure of the data manifold and discriminability in exposing class subspace information, Liu et al. [11] designed a model to integrate locality priors and class information into a unified framework for FSR. Unlike SISR, the reference-based super resolution approach no longer requires only information about the face itself, but also depends on LR with a given high-quality reference facial image. For instance, GFRNet [12] and GWAInet [13] independently proposed to divide the method

into two parts, first aligning the reference image and LR with different poses and expressions for content correction, and then using the reconstruction network RecNet to guide the fusion of the corrected reference image and LR features for FSR. With multiple exemplars, Wang et al. [14] designed a model via the combination of effective features on all reference surfaces by a weighted pixel averaging module, and then added the results to a super-resolution model for high-frequency feature fusion at different scales. There are also some works that use dictionary learning-based methods. Hao et al. [15] proposed a novel FSR reconstruction method based on non-local similarity and multi-scale linear combination consistency. In [16], Ikram et al. designed a reference patch embedding to improve the structural similarity in the LR manifold and adopted a regional constraint to refine the selection of neighbors to better optimize the reconstruction weights. Recently, Li et al. [17] proposed a deep face dictionary network (DFDNet) to reconstruct individual components using a similar dictionary library of different components.

II. METHODOLOGY

As shown in Fig. 1, a multi-source references based face super-resolution network (MSRNet) is proposed. It mainly consists of two parts. We first construct an "appearance-alike" high-quality reference dataset "CelebA_Ref", and then propose MSRNet and execute it in two stages, i.e., external reference mining and internal self-mining, with each sub-part achieving LR feature enhancement and FSR reconstruction by exploring spatial similarity information at the semantic patch level of faces in the reference. Note that the relationship between the two parts is complementary. MSRNet can be formulated as follows:

$$I_{SR} = MSRNet(I_{LR}, I_{Ref}^{m}; \Theta) \tag{1}$$

where I_{SR}, I_{LR}, I_{Ref}^{m} denotes the super-resolution result, the low-resolution input image and the m reference images corresponding to LR, respectively. Θ represents the parameters that can be learned in MSRNet.

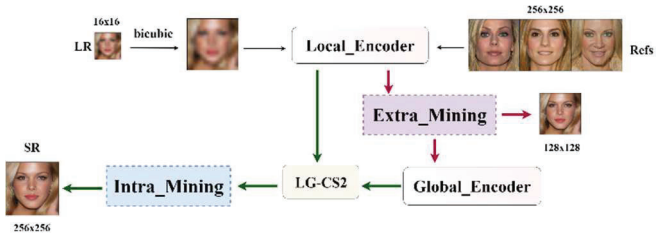


Fig. 1. The overview of MSRNet.

A. Construction of CelebA-Ref.

In this work, CelebA-HQ is used to construct the reference pair, which is a high-quality version of CelebA [18] and consists of 30,000 images with a 1024 × 1024 resolution. Specifically, in the first step, face detection is performed using MTCNN [] and the detected face is then resized to 256 ×256 according to the bounding box. This aims to reduce the interference of different complex backgrounds on face content reconstruction. The filtered and pre-processed dataset is denoted as $\mathbb{Z}$. Then, the pre-trained VGGFace [] network is used to extract multi-layer features F^{l} with different representational capabilities for face images, and the values of l are set to 5, 17 and 29 in the experiment, empirically.

Finally, reference pairs are established. Suppose the current face image is Q, calculate the $\mathcal{L}_1$ distance between it and all the non-identities with it in $\mathbb{Z}$ on each layer of F^{l}. Sum the $\mathcal{L}_1$ distances of each layer as the overall similarity measurement and take the top m items as the reference data of Q after forward sorting. Mathematically, this process can be formulated as follows:

$$D_{un\_Q} = \sum_{i}^{l} \left(\left| F_{Q}^{i} - F_{un_Q}^{i} \right|_1 \right), \tag{2}$$

$$I_{Q}^{m} = Top(D_{un\_Q})_m, \tag{3}$$

where $Top(.)_m$ denotes the operation of taking the top m most similar reference images in D_{un_Q} of I_Q.

B. The designed MSRNet

As is illustrated in Fig. 1, the designed MSRNet mainly consists of two parts: encoders and information mining. The encoders are divided into two forms: CNN-based local feature extractor (LFE) and SwinTransformer Encoder-based global feature extractor (GFE). Information mining part consists of two modules: external mining (EM) based on non-identity references and self-similarity mining (SSM).

LFE with three sets of residual blocks is first deployed to map LR and its corresponding references to the same feature space, denoted as F_{LR}^{s}, $F_{Ref}^{m,s}$, where $s = 1,2,4$, representing features of different scales. Then, the EM module matches and fuses the valuable similar information in the feature space of the multi-source references to generate I_{SR}^{1}. This completes the first stage of the reconstruction task with scale factor of 8. Next, the GFE is used to extract the global long-range correlation of I_{SR}^{1}. The SSM module computes the self-similarity in the process of interacting local and global information across different scales and completes the second up-sampling stage with a magnification factor of 2.

Stage 1: Extra-Mining. It is well known that the content and structure of face images naturally have non-local similarity and symmetry [3], such as left/right eyes, nose, eyebrows, upper/lower lips, etc. However, such deterministic semantic components are obviously limited by factors such as expression, pose and color, thus there is a high possibility that similar semantic regions belong to non-homologous sources. Inspired by MASA-SR [19], to avoid the potential negative impact of the reconstruction results from one single reference, we extend it to multiple external reference sources to capture more accurate reference details in an adaptive patch-level matching and reorganization manner, as follows:

$$r_{c,j}^{m,k} = \langle \frac{p_c^k}{\|p_c^k\|}, \frac{q_j^m}{\|q_j^m\|} \rangle, \tag{4}$$

where p_c^k represents the center patch of the current block among the k non-overlapping blocks of LR, q_j^m denotes the j_th patch of the j_th reference, and $r_{c,j}^{m,k}$ indicates the cosine similarity of the two parts. Note that the operating platform here is the feature map with $s = 4$ in F_{LR}^{s} and $F_{Ref}^{m,s}$. We continue to adopt the coarse-to-fine matching mechanism proposed in MASA. It returns the index of each reference patch, which is then mapped to the feature maps of the other two scales and goes through corresponding cropping and reorganization manipulation. In detail, we propose a cross-scale dual-residual fusion mechanism (CDFM) for multi-source feature fusion and reconstruction. The detailed process is expressed as follows:

$$\begin{cases} F_{Ref}^{res} = Conv\left(Up\left(F_{Ref}^{hi}\right)\right) - F_{LR}^{low} \\ F'_{Ref} = F_{Ref}^{hi} + Deconv(F_{Ref}^{res}) \end{cases}, \tag{5}$$

$$\begin{cases} F_{LR}^{res} = F_{LR}^{hi} - Conv\left(Down\left(F_{Ref}^{low}\right)\right) \\ F'_{LR} = Conv(F_{LR}^{res} + F_{Ref}^{hi}) \end{cases}, \tag{6}$$

$$F_{LR}^{enh} = Conv(cat(F'_{Ref}, F'_{LR})), \tag{7}$$

where F'_{Ref} and F'_{LR} denote the highlighted high-frequency detail information in LR and Ref, respectively. The superscript hi and low express features at different levels, accordingly. Up and $Down$ mean the bicubic interpolation function. This cross-scale fusion of residual features without additional face prior knowledge can not only effectively improve the discriminative power of the network, but also perceive finer visual details by focusing on the specific information of different sources.

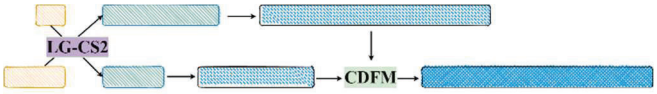


Fig. 2. The pipeline of Stage2.

Stage 2: Intra-Mining. The innate structural symmetry of the human face makes it distinctly non-local in its similarity and dependency. Establishing similar feature dependence across scales and feature spaces allows us to perceive fine detail information and capture more faithful high-frequency details from the heterogeneous spaces, with the purpose of obtaining more accurate reconstruction details.

We designed interactive modules for local and global information cross-scales-cross-spaces module LG-CS2, as shown in Fig.2, where the gradient gray and blue blocks on the left side represent local features F_{LR}^{s} and global features F_{SR}^{g}, respectively. And their lengths denote the size of the feature map. It is known that SwinTransformer [20] solves the problem of information exchange between different windows via introducing shifted window partitions. Therefore, we first use SwinTransformer's encoder to extract the global features of the output of the first stage. Suppose the original size is $(\mathrm{W} \times H)$, re-scale it with a downsampling factor of 2 to produce the feature map of size $(\frac{W}{2} \times \frac{H}{2})$. They construct cross-scale and cross-space feature reference pairs with the intermediate and high-level features in F_{LR}^{s}, respectively. In this manner, LG-CS2 computes the weight of each patch according to Eq. 8 during the interaction between local and global features, while excavating their differences in spatial dimensionality and information representation, and these maps F_{SR}^{g} and $F_{SR}^{g} \downarrow$ are resized into twice the current dimension, i.e., $(2W \times 2H)$ and $(W \times H)$, respectively. Then, we feed these two results into the proposed CDFM module to enhance the fusion of multi-scale information and obtain the final reconstruction outputs:

$$w = \frac{exp\phi(X_{i,j}^{r*r}, Y_{g,h}^{r*r})}{\sum_{u,v} exp\phi(X_{i,j}^{r*r}, Y_{u,v}^{r*r})}, \tag{8}$$

where X and Y represent the feature maps at different scales in the reference pair (i,j) and (g,h), respectively. And (u,v) denotes the coordinates of feature maps, and $\phi(.)$ represents the function used to measure the similarity, according.

III. Experiments

Extensive experiments are performed on an Intel Core i7-11700KF CPU @ 3.60GHz, NVIDIA GeForce RTX3090 GPU, and Ubuntu 20.04 64bit. We use Python 3.7 and Pytorch 1.8.0 to realize MSRNet. The Adam optimizer is adopted to train our model with learning rate $lr = 2 \times 10^{-4}$, $\beta_1 = 0.9$ and $\beta_2 = 0.999$, experimentally.

As a preprocessing step in the experimental part, we expand the input pixel intensity values from [0; 255] to [0; 1]. All intermediate features have $C = 64$ channels. In order to quantify the performance of the individual method and to make a comprehensive comparison with the-state-of-the-art models, PSNR and SSIM are used as quality assessment metrics.

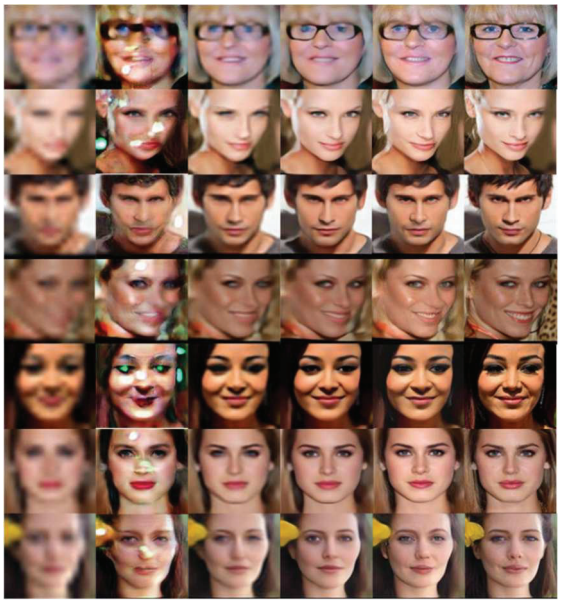
After an initial screening of the data, we selected 18,000 images from Celeba-HQ to build the reference library Celeba-Ref. Among these reference pairs, we randomly selected 14,000 as the training set, 3,000 as the validation set, and the remaining 1,000 is the test set, with no identity duplication between the three datasets. In each reference pair, there are five high-quality reference images of different identities corresponding to the LR input. We employ hyperbolic interpolation function as a degenerate model to synthesize the low-quality training data, and the size in the experiment is set to 16×16.

To verify the effectiveness of the proposed method, we compared it with the following typical methods: Bicubic [21], FSRGAN [7], SwinIR [22], and SPARNet [23]. Maintaining the fairness of the comparison, we retrained all methods in a uniform environment with 50 epochs using the same training data, and the performance of the test dataset at three different sizes is presented separately in Table 1.

TABLE 1. COMPARISON WITH STATE-OF-THE-ART METHODS

Models	x4		x8	
	PSNR	SSIM	PSNR	SSIM
FSRGAN	16.3976	0.5992	17.4598	0.5294
SwinIR	24.9200	0.8577	23.8600	0.7521
SPARNet	27.3597	0.8265	24.8817	0.6779
Ours	**32.5969**	**0.9053**	**30.0643**	**0.8030**

As can be seen from the Table 1, the proposed FSR model performs well on most of the evaluation metrics, especially with the magnification factor of 4 and 8. For qualitative comparison, some representative results for the corresponding tasks are shown in Figure 3. Our MSRNet can complement richer and more convincing visual details, such as glasses, beards, and wrinkles. In contrast, other methods either produce over-smoothed and color-shifted results or generate systematic artifacts when supplementing details. From the viewpoint of visual appearances, MSRNet is capable of reconstructing fine face details in a uniform external reference-based manner, which also verifies the effectiveness of the proposed method.



(a) (b) (c) (d) (e) (f)

Fig. 3. Qualitative results (8x) of our MSRNet and related methods, (a) Bicubic, (b) FSRNet, (c) SwinIR, (d) SPARNet, (e) Ours, (f) Ground truth.

Ablation study is implemented as given in Tables 2 and 3. In detail, Table 2 and Table 3 display the choice of the number of references and the role of the GFE module, respectively. In Table 2, we experimented with the number of references from 1 to 5 for the up-sampling factor of 8 It indicates that the multi-source external references in the first phase is effective and helpful for FSR, while on the other side, it is more critical the quality rather than the quantity of the reference sources. In Table 3, where "no_sw" corresponds to the operation of using a regular 3×3 convolution layer to connect the two stages instead of the self-attended Encoder in SwinTransformer. We can also conclude from the evaluation metrics about the necessity of perceiving facial global information.

TABLE 2. THE ABLATION OF THE NUMBER OF EXTERNAL REFERENCES.

Num.Ref	x8	
	PSNR	SSIM
Ref=1	24.9270	0.7375
Ref=2	29.8756	0.7980
Ref=3	24.2813	0.7176
Ref=4	**30.0643**	**0.8030**
Ref=5	29.8457	0.7924

TABLE 3. THE ABLATION OF GFE.

	x16	
	PSNR	SSIM
no_sw	23.1598	0.6148
with_sw	23.5580	0.6291

IV. CONCLUSIONS

This paper proposes a multi-source references based FSR network. A new identity-agnostic multi-reference database is constructed, and a multi-stage feature fusion and reconstruction model is designed for face hallucination. Experimental results show that the reference information matched at each stage of the model is instrumental and effective. Compared with other typical methods, the designed framework can recover satisfactory HR results with more convincing high-frequency details.

ACKNOWLEDGMENT

This work was supported by National Natural Science Foundation of China (NSFC) (61976123, 61601427); Taishan Young Scholars Program of Shandong Province; and Key Development Program for Basic Research of Shandong Province (ZR2020ZD44).

REFERENCES

[1] J. Jiang, C. Wang, X. Liu, "Deep learning-based face super-resolution: A survey," ACM Computing Surveys, vol. 55, no. 1, pp. 1-36, 2021.

[2] M. W. Jian, K. M. Lam, "Simultaneous Hallucination and Recognition of Low-Resolution Faces Based on Singular Value Decomposition," IEEE Trans. on CSVT, vol. 25, no. 11, pp. 1761-1772, Nov. 2015.

[3] M. W. Jian, K. M. Lam, J. Y. Dong, "Facial-Feature Detection and Localization Based on a Hierarchical Scheme," Information Sciences, vol. 262, pp. 1-14, Mar. 2014.

[4] W. Huang, Y. Chen, L. Mei, "Super-resolution reconstruction of face image based on convolution network," in Proc. of 2017 International Conference on Intelligent and Interactive Systems and Applications, vol. 686, pp. 288-294, Nov. 2017.

[5] K. Jiang, Z. Wang, P. Yi, "Dual-path deep fusion network for face image hallucination," IEEE Transactions on Neural Networks and Learning Systems, vol. 33, no. 1, pp. 378 - 391, Oct. 2020.

[6] J. Jiang, J. Ma, and S. Tang, "Face hallucination through differential evolution parameter map learning with facial structure prior," Information Sciences, vol.481, pp. 174-188, May.2019.

[7] Y. Chen, Y. Tai, and X. Liu, "Fsrnet: End-to-end learning face super-resolution with facial priors," in Proc. of 2018 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2492-2501, 2018.

[8] Y. Zhang, Y. Wu, and L. Chen, "MSFSR: A multi-stage face super-resolution with accurate facial representation via enhanced facial boundaries," in Proc. of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pp. 504-505, 28 July. 2020.

[9] L. Liu, C.L.P Chen, and Y. Wang, "Modal Regression-Based Graph Representation for Noise Robust Face Hallucination," IEEE Transactions on Neural Networks and Learning Systems, doi:10.1109/TNNLS.2021.3106773, pp. 1-13, 06 Sep. 2021

[10] L. Chen, J. Pan, and J. Jiang, "Multi-stage degradation homogenization for super-resolution of face images with extreme degradations," IEEE Transactions on Image Processing, 2021.

[11] L. Liu, R. Lan, Y. Wang, "Discriminative face hallucination via locality-constrained and category embedding representation," IEEE Transactions on Systems, Man, and Cybernetics: Systems, vol. 51, no. 12, pp. 7314-7325, 2020.

[12] X. Li, M. Liu, Y. Ye, et al, "Learning warped guidance for blind face restoration," in Proc. of 2018 European conference on computer vision (ECCV), doi: https://doi.org/10.1007/978-3-030-01261-8_17, pp. 278-296, Oct. 2018.

[13] B. Dogan, S. Gu, R. Timofte, "Exemplar guided face image super-resolution without facial landmarks," in Proc. of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pp. 1814-1823, June. 2019.

[14] K. Wang, J. Oramas, T. Tuytelaars, "Multiple exemplars-based hallucination for face super-resolution and editing," in Proc. of 2021 Asian Conference on Computer Vision (ACCV), pp. 258-273, Feb. 2021.

[15] N. Hao, H. Liao, Y. Qiu, et al, "Face super-resolution reconstruction and recognition using non-local similarity dictionary learning based algorithm," IEEE/CAA journal of Automatica Sinica, vol. 3, no. 2, pp. 213-224, Apr. 2016.

[16] J. Ikram, Y. Lu, J. W. Li, N. Hui, H. Bokhari, "Face Hallucination in a high resolution feature space using an intermediate dictionary learning via reference patch embedding," IEEE/CAA journal of Automatica Sinica, pp. 1−12, Dec. 2017.

[17] X. Li, C. Chen, S. Zhou, et al. "Blind face restoration via deep multi-scale component dictionaries," in Proc. of 2020 European Conference on Computer Vision (ECCV), pp. 399-415. Nov. 2020.

[18] Z. Liu, P. Luo, X. Wang, et al. “Deep learning face attributes in the wild,” in Proc. of 2015 IEEE international conference on computer vision (ICCV), pp. 3730-3738. Dec. 2015

[19] L. Lu, W. Li, X .Tao, et al. “Masa-sr: Matching acceleration and spatial adaptation for reference-based image super-resolution” in Proc. of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), doi: 10.1109/CVPR46437.2021.00630, pp. 6368-6377. Jun. 2021.

[20] Z. Liu, Y. Lin, Y. Cao, et al. “Swin transformer: Hierarchical vision transformer using shifted windows” in Proc. of 2021 IEEE/CVF International Conference on Computer Vision (ICCV), pp. 10012-10022. Oct. 2021.

[21] H. S. Hou, H. C. Andrews, "Cubic splines for image interpolation and digital filtering," IEEE Transactions on Acoustics, Speech, and Signal Processing, vol. 26, no. 6, pp. 508-517, Dec. 1978.

[22] J. Liang, J. Cao, G. Sun, et al. “Swinir: Image restoration using swin transformer,” in Proc. of 2021 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW). pp. 1833-1844. Oct. 2021.

[23] C. Chen, D. Gong, H. Wang, et al. Learning spatial attention for face super-resolution. IEEE Transactions on Image Processing, vol. 30, pp. 1219-1231, Dec. 2020.

[24] M. Jian, C. Cui, X. Nie, H. Zhang, L. Nie, Y.Yin, Multi-view face hallucination using SVD and a mapping model. Inf. Sci. 488: 181-189, 2019.

[25] M. Jian, K. M. Lam, Face-image retrieval based on singular values and potential-field representation. Signal Process. 100: 9-15, 2014.

[26] M. Jian, K.M. Lam, J. Dong, A novel face-hallucination scheme based on singular value decomposition. Pattern Recognit. 46(11): 3091-3102, 2013.