IEICE TRANSACTIONS

on Fundamentals of Electronics, Communications and Computer Sciences

DOI:10.1587/transfun.2024EAL2027

Publicized:2024/07/16

A PUBLICATION OF THE ENGINEERING SCIENCES SOCIETY

The Institute of Electronics, Information and Communication Engineers

Kikai-Shinko-Kaikan Bldg., 5-8, Shibakoen 3 chome, Minato-ku, TOKYO, 105-0011 JAPAN

LETTER

Noisy face super-resolution method based on three-level information representation constraints

Qi QI[†a)], ***Nonmember***, **Zi TENG**[††], ***Nonmember***, **Hongmei HUO**[†††], ***Nonmember***, **Ming XU**[††††], ***Member***, **Bing BAI**[†††††], ***Nonmember***

SUMMARY To super-resolve low-resolution (LR) face image suffering from strong noise and fuzzy interference, we present a novel approach for noisy face super-resolution (SR) that is based on three-level information representation constraints. To begin with, we develop a feature distillation network that focuses on extracting pertinent face information, which incorporates both statistical anti-interference models and latent contrast algorithms. Subsequently, we incorporate a face identity embedding model and a discrete wavelet transform model, which serve as additional supervision mechanisms for the reconstruction process. The face identity embedding model ensures the reconstruction of identity information in hypersphere identity metric space, while the discrete wavelet transform model operates in the wavelet domain to supervise the restoration of spatial structures. The experimental results clearly demonstrate the efficacy of our proposed method, which is evident through the lower Learned Perceptual Image Patch Similarity (LPIPS) score and Fréchet Inception Distances (FID), and overall practicability of the reconstructed images.

key words: feature distillation, latent information constraint, image super-resolution, image denoising, deep neural network, hypersphere metric space, wavelet transform.

1. Introduction

Face SR is an important research branch in the field of computer vision, mainly aimed at reconstructing high-resolution (HR) face image from one or more LR face images captured in the same scene. Early research has focused on interpolation methods [1], neighborhood embedding [2], and sparse representation [3]. Although these algorithms can achieve certain results, problems such as ringing, blurring, and artifact often occur in reconstructed images, severely limiting the quality of reconstructed images.

As deep learning methods gradually become the mainstream research direction in the field of computer vision, the successful application of deep neural networks in feature extraction and nonlinear mapping provides new solutions for solving the problem of face SR. For example, Cao et al. proposed a face SR algorithm based on attention perception mechanism [4], which uses deep reinforcement learning to find face patches participating in face SR. Ma et al. proposed an iterative collaborative convolutional neural network that utilizes two recurrent neural networks for face image reconstruction and key point prediction [5]. Wang et al. proposed applying generative facial priors to face SR problems and proposed a channel splitting spatial feature transformation model to fuse reconstruction features with prior features [6]. However, due to the fact that generative prior based methods can only recover face images with a fixed style (FFHQ [7] dataset style), in order to improve the effect of human visual perception, detail features unrelated to the original face are generated, and the generated results cannot effectively preserve identity information. In addition, existing face SR methods are conducted under the assumption that the input image is noise free. In practical application scenarios, input images contaminated by noise can lead to a sharp decline in model performance, and reconstructed face images may have obvious identity information confusion, which cannot meet practical application needs.

To solve the problem of existing methods being unable to effectively remove the interference of degradation factors such as strong noise and blur on the face reconstruction process, we propose a noisy face SR method based on three-level information representation constraints. (1) We design a feature distillation network to extract effective face information, which exploits statistical anti-interference model and latent contrast algorithm to remove invalid information. (2) We design a face reconstruction network, which utilizes the extracted face features to reconstruct HR face images. (3) We deploy a face identity embedding model and discrete wavelet transform model to further supervise the reconstruction of identity information and spatial structure from the hypersphere identity metric space and wavelet domain respectively.

2. Proposed Method

2.1 Network Architecture

As shown in Fig.1, the proposed network mainly consists of four parts, namely feature distillation network, face reconstruction network, identity information embedding module, and discrete wavelet transform module, which is a three-level framework of feature level reconstruction, semantic level reconstruction, and pixel level reconstruction. (1) Firstly, the feature distillation network removes invalid information such as noise from face images by designing statistical anti-interference models (SAIB) and latent space

[†]The author is with the Department of Decision Consulting, Party School of Liaoning Provincial Party Committee, Shenyang, 110004, China.
[††]The author is with the School of Robotics Science and Engineering, Northeastern University, Shenyang, 110819, China.
[†††]The author is with the Department of Decision Consulting, Party School of Liaoning Provincial Party Committee, Shenyang, 110004, China.
[††††]The author is with the School of Artificial Intelligence, Shenyang University of Technology, Shenyang, 110870, China.
[†††††]The author is with the Department of Leadership Science, Party School of Liaoning Provincial Party Committee, Shenyang, 110004, China.
a) E-mail: qiqi@stu.syau.edu.cn

feature comparison algorithms, thereby achieving high robustness in latent space feature level reconstruction. (2) Then, the identity information embedding module applies identity recognition constraints to reduce the identity difference between the reconstructed face image and the HR face image, achieving semantic level reconstruction of the face image. (3) Finally, the discrete wavelet transform module captures the spatial and frequency information, decouples the reconstructed face image and HR face into different frequency sub bands, and calculates the corresponding high-frequency band loss to achieve pixel level reconstruction of face high-frequency details. The feature distillation network and the face reconstruction network are composed of multiple residual attention blocks (RCAB) [8] and Transformer blocks [9]. The corresponding layers of the two sub networks are connected across layers to maximize the information flow between the convolution layers.

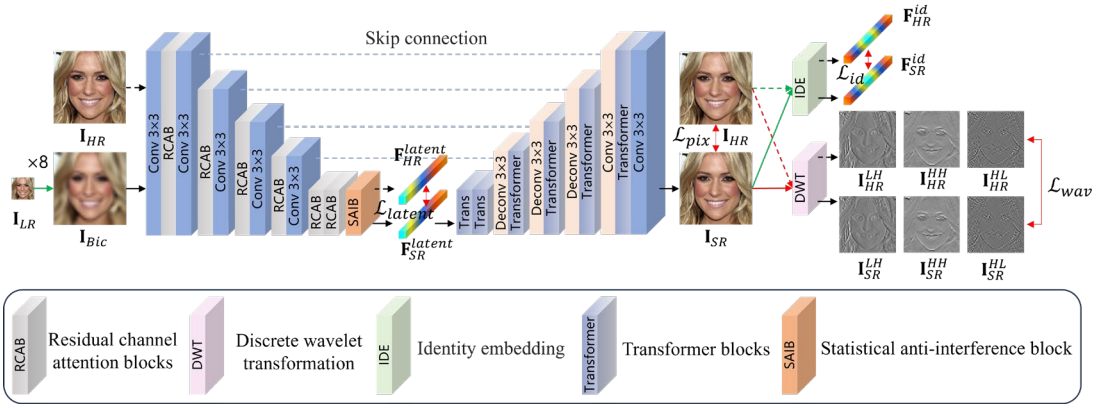

Fig. 1 Illustration of the proposed face SR network.

2.2 Latent Space Feature Distillation

Existing methods mainly constrain the reconstruction process of face information from the pixel level. However, in noisy LR face images, accurate reconstruction of pixel level information is very difficult, and simply relying on pixel level information constraints can easily lead to semantic distortion. To address this issue, our method directly calculates the loss in the latent space and constrains the latent features, effectively improving the reconstruction accuracy of latent features. At the same time, utilizing latent feature loss for backpropagation in the middle part of the network can better guide the training process of the encoder network and improve the feature extraction ability. The reconstructed latent features can also be better expressed in the decoder network, efficiently guiding the decoder to recover face features.

In order to avoid the interference of noise and fuzzy information while extracting core features in the distillation network, a statistical anti-interference block (SAIB) is used to add random Gaussian noise δ with a mean of 0 and a variance of 1 to the extracted features. Two convolution layers of $g(\cdot)$ are used to process the encoded features, which adds disturbance to improve the anti-interference ability, making the network focus on more critical information and obtain latent features with stronger representation ability. The feature distillation loss $\mathcal{L}_{latent}$ based on SAIB is defined as follow:

$$F_t^{latent} = SAIB(Dis(\theta_1, I_t)) = g(Dis(\theta_1, I_t) + \delta), \quad (1)$$

$$\mathcal{L}_{latent}^{i} = \|F_{HR}^{latent} - F_{SR}^{latent}\|_1 \quad (2)$$

where θ_1 is the parameter of the distillation network $Dis(\cdot)$. F_t^{latent} denotes the latent feature, which is obtained by the distillation feature $Dis(\theta_1, I_t)$ and statistical anti-interference module $SAIB(\cdot)$. t represents that the input image I belongs to either HR or noisy LR faces.

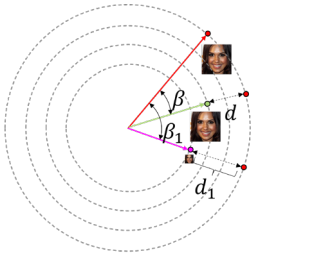

Fig. 2 hypersphere metric space.

2.3 Face Identity Embedding

As a specific application problem in the field of image SR, face SR can be guided by facial prior information in the reconstruction process, such as face component heatmap [10], face shape with key points [11], face wavelet coefficients [12], and face local structural prior information [13]. However, in high noise environments, the semantic prior information in noisy LR face image is difficult to be accurately estimated, so the reconstruction performance of these methods will also show a significant decline. Unlike these methods that directly use face semantic prior information, our method uses face identity prior information to guide the face image reconstruction with realistic texture details and unchanged identity information.

In order to accurately recover identity related face details, we propose an identity recognition constraint model to reduce the identity difference between reconstructed face

images and HR face images. Firstly, due to the excellent performance of hypersphere space in face identity representation, we use it as an identity metric space and utilize the pretrained identity information embedding model LightCNNv9 [14] to extract identity related features. Then, Euclidean regularization is used to map identity features to a hypersphere space for identity loss calculating.

As shown in Fig.2, there are significant differences in angle and amplitude between noisy LR face features and denoised HR face features in the hypersphere identity metric space. And with the decrease in resolution, the average angle and amplitude differences between low- and high-resolution face feature pairs will gradually increase. This indicates that the two indicators of angle and amplitude can reflect the degree of representation features degradation, and use them to predict the quality of face reconstruction images. Therefore, we propose a feature deconstruction based identity recognition loss, which calculates the loss of reconstructed face identity vectors and HR face identity vectors in terms of angle and amplitude in the hypersphere metric space ($\mathcal{L}_a$,$\mathcal{L}_m$), and adds them up to obtain identity recognition loss $\mathcal{L}_{id}$:

$$\mathcal{L}_a = 1 - \frac{(F_{SR}^{id})^T (F_{HR}^{id})}{\|F_{SR}^{id}\|_2 \|F_{HR}^{id}\|_2}, \tag{3}$$

$$\mathcal{L}_m = \| \|F_{SR}^{id}\|_2 - \|F_{HR}^{id}\|_2 \|_2, \tag{4}$$

$$\mathcal{L}_{id} = \mathcal{L}_a + \mathcal{L}_m \tag{5}$$

where F_{SR}^{id} and F_{HR}^{id} represent the identity feature vectors produced by the identity information embedding model LightCNNv9. By precisely constraining the reconstructed face identity vector in terms of angle and amplitude in the hypersphere metric space, our model can reconstruct HR face images with realistic texture details and unchanged identity information.

2.4 Wavelet Knowledge Distillation

Existing methods can accurately reconstruct low-frequency details of face images, but they often cannot effectively reconstruct high-frequency details, and this phenomenon is more pronounced in high noise environments. To solve this problem, we propose to use the discrete wavelet transform method to decouple the reconstructed face image and HR reconstructed face into sub bands of different frequencies. Then we calculate the L1 loss corresponding to the high-frequency band to constrain the reconstruction process of high-frequency face details.

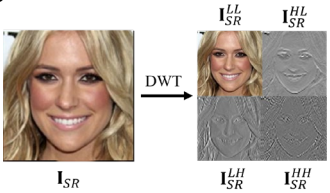

Fig. 3 discrete wavelet transform.

The discrete wavelet transform method is commonly used as a mathematical tool for decoupling pyramid images. In our method, reconstructed face images and HR face images are decoupled into four sub bands, namely LL, LH, HL, HH, where LL represents the low-frequency sub band and the rest are the high-frequency sub bands. As shown in Fig.3, if the discrete wavelet transform is defined as $\Psi(\cdot)$, the high and low sub band images of image I can be represented as $\Psi^H(\cdot)$ and $\Psi^L(\cdot)$, respectively. Wavelet knowledge loss $\mathcal{L}_{wav}$ is defined as follows:

$$\mathcal{L}_{wav} = \|\Psi^H(I_{HR}) - \Psi^H(I_{SR})\|_1, \ \Psi^H(\cdot) = I^{LH}, I^{HL}, I^{HH} \tag{6}$$

where I_{SR} and I_{HR} represent SR and HR face images, respectively. Compared to other frequency analysis methods such as Fourier transform, wavelet transform can more effectively capture spatial and frequency information in image signals [15]. The effectiveness of the wavelet knowledge distillation in recovering high-frequency details lies in its ability to decompose images into different frequency bands, capturing details at various scales. This decomposition enables the concentration of face SR model in high-frequency bands, where fine textures and details are effectively enhanced. Additionally, the localized processing nature of wavelet transform allows for precise analysis of local features, aiding in the restoration of facial details. Moreover, the reversibility of wavelet transform ensures that adjustments made in the wavelet domain can be accurately applied to reconstruct the original image, enhancing the face super-resolution performace, which enables our model to reconstruct face images with high-quality high-frequency detail textures.

3. Experiments

3.1 Experimental Setup

Based on experimental experience, training with entire CelebA dataset [16] is unable to significantly improve the performance of face super-resolution network, but it will significantly increase the network training time. Therefore, we selected 40000 face images from the CelebA for training and use the next 1000 face images as the testset. To demonstrate that our method can reconstruct clear face images under noise and fuzzy interference, three degradation models were used in the experiment to synthesis LR face images. We use the bicubic operation to produce LR images with a scale factor of 8 (***Bic***). Then, Gaussian noise with a noise level of 15 is added to the 8-fold downsampled image to obtain LR face images with noise (***BicN***). Finally, in order to produce degraded face images that are simultaneously affected by noise and blurring factors, we first applied Gaussian kernel blur with a standard deviation of 1.5 and size of 7×7 to HR faces, then perform bicubic downsampling with a scale factor of 8 on these images, and then add Gaussian noise with a noise level of 30 (***BBicN***).

In terms of training setting, Adam algorithm is used as the loss function optimizer. The batch size is 16. The model was trained on a TITAN X GPU. To evaluate the quality of face SR results, we employ the Learned Perceptual Image

Patch Similarity (LPIPS) score [17] and Fréchet Inception Distances (FID) [18] to assess the perceptual realism of generated faces, as pixel space metrics only measure local distortions and may not align with human perception.

Tab. 1 quantitative evaluation results of our method and SOTA methods.

Method	***Bic***		***BicN***		***BBicN***	
	LPIPS↓	FID↓	LPIPS↓	FID↓	LPIPS↓	FID↓
Panini-Net	55.63	0.4538	81.87	0.5449	90.42	0.5931
DIC	36.88	0.3779	75.95	0.5263	94.8	0.6247
GPEN	60.02	0.471	65.43	0.4811	86.59	0.57
GFPGAN	50.5	0.4295	53.41	0.4312	70.94	0.5031
Ours	30.47	0.3284	43.92	0.382	47.05	0.4062

3.2 Comparative Experiments

In order to evaluate the performance of our method, the most advanced face SR methods PaniniNet [19], DIC [5], GPEN [20], GFPGAN [6] were selected for comparative experiments on CelebA testsets under different degradation processes. Tab.1 presents the quantitative evaluation results of our method and these state of the arts (SOTA) methods on the CelebA dataset. The quantitative comparison results show that the reconstruction results of our method are significantly lower than the current technical level in terms of LPIPS and FID performance. In order to further evaluate the visual effect of our method, a qualitative comparison was conducted under different degradation processes.

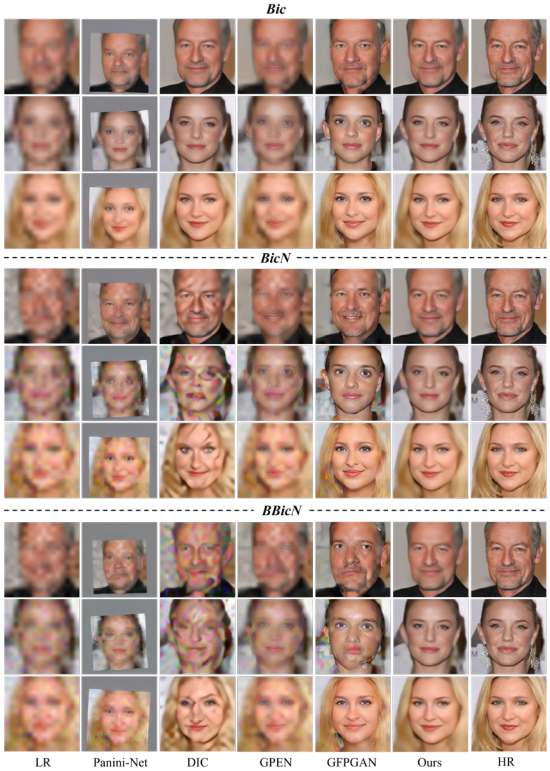

Fig. 4 Qualitative comparison results of our method and existing methods on the all the degradation models.

As shown in Fig.4, in the absence of noise and blur (***Bic***), existing methods can reconstruct ideal facial details while preserving identity information to a certain extent. However, under the influence of noise, the performance of existing methods has shown varying degrees of decline, and our method can still reconstruct clearer face images. As shown in Fig.5, the face image reconstructed by our method has clearer face texture details and can better preserve identity information. Although GFPGAN combines generative facial priors with reconstruction features to reconstruct face images with high-frequency detail textures, it can only recover face images with fixed styles but fail in retaining identity information.

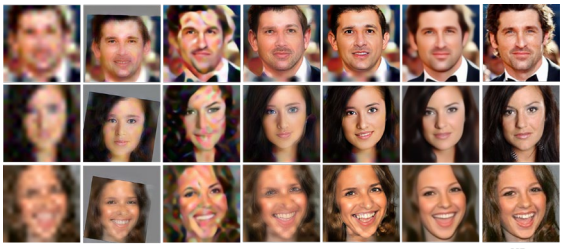

Fig. 5 Qualitative comparison results of our method and existing methods on the ***BicN*** degradation model.

As shown in Fig.6, after introducing the fuzzy degradation factor, the contaminated LR input will cause a sharp decline in the performance of each model, and the reconstructed face image will have obvious identity information confusion, which cannot meet practical application needs. Benefit by the proposed three-level information representation constraints method, it can be clearly seen that although the degradation process is gradually becoming more complex, the visual quality of the face images reconstructed by our model has not significantly decreased, so it has good practicality.

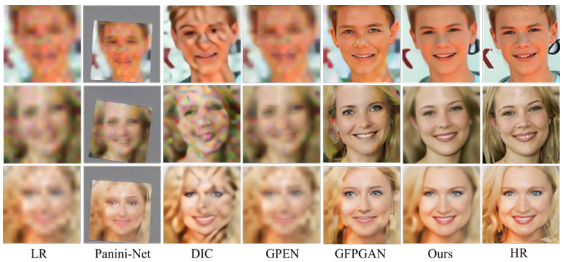

Fig. 6 Qualitative comparison results of our method and existing methods on the ***BBicN*** degradation model.

4. Conclusion

In this paper, we designed a feature distillation network to extract effective facial information, which exploited statistical anti-interference model and latent contrast algorithm to removed invalid information such as noise. And we designed a face reconstruction network, which utilized the extracted face features to reconstruct HR face images. Finally, we deployed a face identity embedding model and discrete wavelet transform model to further supervise the reconstruction of identity information and spatial structure from the hypersphere identity metric space and wavelet domain respectively. The experimental results showed that the proposed method not only removed the noise from face in the high noise environment, but also improved the resolution of the face image effectively, which obtains better LPIPS and FID performance, and good practicability.

References

[1] Fritsch F N, Carlson R E. Monotone Piecewise Cubic Interpolation[J]. Siam Journal on Numerical Analysis, 1980, 17(2): 238-246.

[2] Liu A, Liu Y, Gu J, et al. Blind image super-resolution: A survey and beyond[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2022, 20(12): 2884-2901.

[3] Yang J, Wright J, Huang T S, et al. Image super-resolution via sparse representation[J]. IEEE Transactions on Image Processing, 2010, 19(11): 2861-2873.

[4] Cao Q, Lin L, Shi Y, et al. Attention-aware face hallucination via deep reinforcement learning[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017: 690-698.

[5] Ma C, Jiang Z, Rao Y, et al. Deep face super-resolution with iterative collaboration between attentive recovery and landmark estimation[C]. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020: 5569-5578.

[6] Wang X, Li Y, Zhang H, et al. Towards real-world blind face restoration with generative facial prior[C]. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021: 9168-9178.

[7] Karras T, Laine S, Aila T. A style-based generator architecture for generative adversarial networks[C]. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019: 4401-4410.

[8] Zhang Y, Li K, Li K, et al. Image super-resolution using very deep residual channel attention networks[C]. Proceedings of the European conference on computer vision (ECCV). 2018: 286-301.

[9] Han K, Xiao A, Wu E, et al. Transformer in transformer[J]. Advances in Neural Information Processing Systems, 2021, 34: 15908-15919.

[10] Yu X, Fernando B, Ghanem B, et al. Face super-resolution guided by facial component heatmaps[C]. Proceedings of the European Conference on Computer Vision (ECCV). 2018: 217-233.

[11] Chen Y, Tai Y, Liu X, et al. Fsrnet: End-to-end learning face super-resolution with facial priors[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018: 2492-2501.

[12] GHuang H, He R, Sun Z, et al. Wavelet-srnet: A wavelet-based cnn for multi-scale face super resolution[C]. Proceedings of the IEEE International Conference on Computer Vision. 2017: 1689-1697.

[13] Jiang J, Chen C, Ma J, et al. SRLSP: A face image super-resolution algorithm using smooth regression with local structure prior[J]. IEEE Transactions on Multimedia, 2016, 19(1): 27-40.

[14] Wu X, He R, Sun Z, et al. A light CNN for deep face representation with noisy labels[J]. IEEE Transactions on Information Forensics and Security, 2018, 13(11): 2884-2896.

[15] Mallat S. A wavelet tour of signal processing[M]. Elsevier, 1999: 83-85.

[16] Liu Z, Luo P, Wang X, et al. Deep learning face attributes in the wild[C]. Proceedings of the IEEE International Conference on Computer Vision. 2015: 3730-3738.

[17] Zhang R, Isola P, Efros A A, et al. The unreasonable effectiveness of deep features as a perceptual metric[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018: 586-595.

[18] Heusel M, Ramsauer H, Unterthiner T, et al. Gans trained by a two time-scale update rule converge to a local nash equilibrium[J]. Advances in Neural Information Processing Systems, 2017, 30.

[19] Wang Y, Hu Y, Zhang J. Panini-net: Gan prior based degradation-aware feature interpolation for face restoration[C]. Proceedings of the AAAI Conference on Artificial Intelligence. 2022, 36(3): 2576-2584.

[20] Yang T, Ren P, Xie X, et al. Gan prior embedded network for blind face restoration in the wild[C]. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021: 672-681.