LETTER

A Dual-Branch Algorithm for Semantic-Focused Face Super-Resolution Reconstruction

Qi QI[†], **Liuyi MENG**[††a)], ***Nonmembers***, **Ming XU**[†††], ***Member***, ***and*** **Bing BAI**[††††], ***Nonmember***

SUMMARY In face super-resolution reconstruction, the interference caused by the texture and color of the hair region on the details and contours of the face region can negatively affect the reconstruction results. This paper proposes a semantic-based, dual-branch face super-resolution algorithm to address the issue of varying reconstruction complexities and mutual interference among different pixel semantics in face images. The algorithm clusters pixel semantic data to create a hierarchical representation, distinguishing between facial pixel regions and hair pixel regions. Subsequently, independent image enhancement is applied to these distinct pixel regions to mitigate their interference, resulting in a vivid, super-resolution face image.

key words: *semantic information, dual branch, image layering, face super-resolution*

1. Introduction

Face super-resolution (FSR) is a specialized image super-resolution (SR) technique that focuses on recovering high-resolution (HR) face images from low-resolution (LR) face images. The human face is a highly structured object with distinctive characteristics that are valuable for the task of FSR. These specific attributes and structure of the human face, such as facial landmarks, skin texture, and facial symmetry, can be effectively explored and leveraged in FSR algorithms. By incorporating this knowledge into the FSR process, it is possible to enhance the quality and resolution of face images more accurately.

In the early years of FSR researches, methods like SPARNet [1] utilized the extracted face landmarks to guide the FSR process. However, accurately extracting landmarks from LR face images can be challenging. Recently, with the rapid development of GAN techniques [2], generative priors of pretrained face GAN models, such as PULSE [3], DIC-GAN [4], GFPGAN [5] are exploited for real-world FSR. These methods first map the LR input image to an intermediate latent code, which then controls the pretrained GAN at each convolution layer to provide generative priors such as facial textures and colors. However, when the given face images have tiny resolution (e.g., of size 16×16 pixels) and arbitrary characteristics which need to be reconstructed at high magnification factors (e.g., 8×), such a decoupling control method is insufficient to guide the precise SR process and leads to unstable quality of restored faces.

Natural images have countless pixel semantics that cannot be layered due to their variability. However, the face image, which consists of the face and a variable background, can be layered. This is because the FSR algorithm mainly focuses on reconstructing the face, allowing the background pixels to be categorized. The face structure and pixel semantics in the image are fixed. Therefore, pixel layering can be performed based on the semantic information, thus easing the reconstruction of different regions. Our analysis revealed the presence of similar image blocks within the face image. Subsequently, we clustered pixels based on semantic similarities. For the purpose of this paper, we defined two categories: the facial pixel region and the hair pixel region, as depicted in Fig. 1.

Consequently, we propose a semantic-based two-branch FSR algorithm. Firstly, the resolution of the LR face image is enhanced using the residual super-resolution module (SRResNet) [6]. Then, the semantics of the pixels in the enhanced face image are identified by using the BiSeNet [7], which results in a hierarchical image. Then, the hierarchical image is fed into the proposed dual-branch structure network. This structure serves two purposes: 1) Further enhancing the initial SR face image; 2) Reducing the interference among different facial semantics regions which have varying reconstruction difficulties. Finally, we design a novel structure extraction module and a self-enhancing module to further enhance the high frequency detail information and improve the overall reconstruction quality. Extensive experiments are implemented on face datasets: CelebA [8], Helen [9], and FFHQ [10], which fully demonstrates that our proposed method can match and exceed the state-of-the-art performance in both quantitative and qualitative measurements.

Fig. 1 The image of facial semantic clustering segmentation.

Manuscript received January 12, 2024.
Manuscript revised February 20, 2024.
Manuscript publicized March 18, 2024.

[†]Department of Decision Consulting, Party School of Liaoning Provincial Party Committee, Shenyang, 110004, China.
[††]School of Robotics Science and Engineering, Northeastern University, Shenyang, 110819, China.
[†††]School of Artificial Intelligence, Shenyang University of Technology, Shenyang, 110870, China.
[††††]Department of Leadership Science, Party School of Liaoning Provincial Party Committee, Shenyang, 110004, China.
a) E-mail: mengliuyi_neu@126.com

DOI: 10.1587/transfun.2024EAL2004

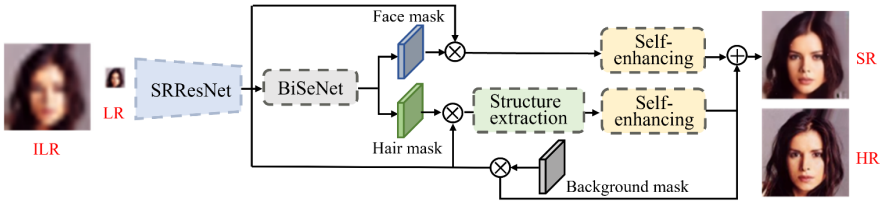

Fig. 2 Illustration of the proposed SR network. ⊗ denotes element-wise multiplication.

2. Proposed Method

2.1 Network Architecture

The network structure of the semantic-focused dual-branch FSR algorithm is composed of three main modules: SR residual module (SRResNet) [6], image layering module (BiSeNet) [7], and the dual-branch high-frequency information enhancement module. The dual-branch module consists of a hair-region enhancement module and a face-region enhancement module. The overall network structure can be seen in Fig. 2.

Face semantic segmentation models, like BiSeNet, tend to have low accuracy when applied to LR face images. To address this, firstly, we improve the effective resolution of the input LR face image with a SR residual module (SRResNet), which can effectively improve the accuracy of subsequent facial segmentation results, as illustrated in Eq. (1).

$$I_{\text{PreSR}} = \text{SRResNet}\,(I_{\text{LR}}) \tag{1}$$

where I_{PreSR} is the initial SR face image and I_{LR} is the observed LR face image. SRResNet can improve the effective face image resolution from 16×16 to 128×128. The enhancement procedure involves the restoration of missing information in the LR face image.

Then, the image layering stage begins by using BiSeNet to extract the semantic information from pixels in the I_{PreSR} image, producing a facial region mask (F_{mask}) and a hair region mask (H_{mask}) which are used to segment the entire face image I_{PreSR} and are non-overlapping, as illustrated in Eq. (2).

$$\begin{aligned} I_{\text{PreSR\_F}} &= \text{F}_{\text{mask}} \times I_{\text{PreSR}} \\ I_{\text{PreSR\_H}} &= \text{H}_{\text{mask}} \times I_{\text{PreSR}} \end{aligned} \tag{2}$$

2.2 Structure Extraction Module

Conventional FSR methods like SPARNet [1] are effective in reconstructing the flat regions in HR face images. However, when the given face images have tiny resolution, the high-frequency textures restored by these methods are either blurred or distorted, which is mainly due to the insufficient attention and supervision for the structure information. To address this issue, a structure extraction module is proposed to restore the high-frequency structure information. The proposed structure extraction module has the convolution layer structure, which is composed of 9 cosine angles

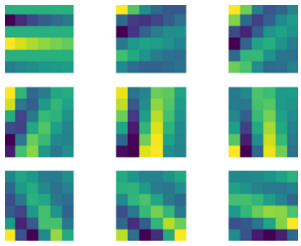

Fig. 3 Structure extraction convolution layer consisting of 9 cosine Gabor kernels.

Gabor kernels [11], denoted as K_θ. As shown in Fig. 3, the structure extraction module has 9 different directional filters, which can effectively capture the intricate details and further restore high-frequency structure information.

Given the hair-region input image, the structure extraction module can generate a set of orientation-specific responses, represented as F_θ, as detailed in Eq. (3).

$$F_\theta = I_{\text{PreSR\_H}} * K_\theta \tag{3}$$

where $*$ represents the convolution operation and the parameters of the convolution kernel are fixed constants. The direction angle θ is sampled uniformly between 0 and π (radian system), and the parameters of the Gabor kernel are calculated as specified in Eq. (4). This choice of Gabor kernel allows for effective extraction of high-frequency structure and contributes to enhancing the quality of the reconstructed image.

$$\begin{aligned} K_\theta &= \exp\left(-\frac{1}{2}\left(\frac{\hat{u}^2}{\sigma_u^2} + \frac{\hat{v}^2}{\sigma_v^2}\right)\right)\cos\left(\frac{2\pi\hat{u}}{\lambda}\right) \\ \hat{u} &= u\cos\theta + v\sin\theta \end{aligned} \tag{4}$$

where σ_u, σ_v and λ are hyperparameters. To find the best performance of our proposed method, we investigate the correlation of these hyperparameters on CelebA dataset. From σ_u, σ_v and $\lambda = 1$ to 5, the average PSNR of different combinations help us determine the selection of hyperparameters, which are set to 1.8, 2.4 and 4 respectively.

To obtain the haired texture map $\boldsymbol{T}(\boldsymbol{i},\boldsymbol{j})$ and orientation map $\boldsymbol{P}(\boldsymbol{i},\boldsymbol{j})$, the maximum value in the feature vector $\boldsymbol{F}(\boldsymbol{i},\boldsymbol{j})$ and its corresponding θ are determined pixel by pixel. This process is described in Eq. (5).

$$\begin{aligned} \boldsymbol{T}(\boldsymbol{i},\boldsymbol{j}) &= \max_\theta \boldsymbol{F}(\boldsymbol{i},\boldsymbol{j}) \\ \boldsymbol{P}(\boldsymbol{i},\boldsymbol{j}) &= arg\,\max_\theta \boldsymbol{F}(\boldsymbol{i},\boldsymbol{j}) \end{aligned} \tag{5}$$

where the functions $\max_\theta(\ )$ and $\text{argmax}_\theta(\ )$ correspond to

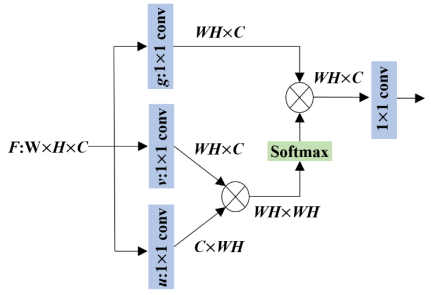

Fig. 4 Self-enhancing module. ⊗ denotes matrix multiplication.

taking the maximum value and its corresponding angle θ in the feature vector $\boldsymbol{F}(\boldsymbol{i}, \boldsymbol{j})$, respectively. After that, $\boldsymbol{T}$ and $\boldsymbol{P}$ are concatenated and sent to the self-enhancing module.

By accurately activating the essential high-frequency structure information with our convolution kernel parameter modulation method, our proposed structure extraction enhancement module can capture the intricate details explicitly and efficiently, and further improves the overall reconstruction quality.

2.3 Self-Enhancing Module

The proposed self-enhancing module exploits the self-similarity measurements and inter spatial correlation information to enhance the features representation. As shown in Fig. 4, inspired by [12], self-enhancing module exploits two streams 1×1 convolutions (u( ) and v( )) to obtain the feature representation, and multiply them to get the response matrix.

Softmax( ) denotes the normalization operation. Enhanced feature is obtained by multiplying response weight matrix and feature representation computed by 1×1 convolutions g( ). The response matrix establishes the long-distance dependence between different locations, which can effectively expand the receptive fields of our network. Thus it overcomes the defects that conventional methods can only obtain the limited information from the local neighbor region, and enhance the high-frequency components and learn abstract feature representations in self-enhancing way.

2.4 Loss Function

BiSeNet is trained to produce the mask maps M_{out} that approximate the ground-truth mask maps M_{gt}, as described in Eq. (6).

$$L_{BiSeNet} = \frac{1}{N}\sum_{i=1}^{N} CrossEntropy\left(M_{out}, M_{gt}\right) \tag{6}$$

where *CrossEntropy*( ) stands for the CrossEntropy·loss·[7].

In the calculation of the reconstruction loss function, all loss terms comply with the constraints of the L_1 paradigm, as described in Eq. (7).

$$L_{pixel} = \frac{1}{N}\sum_{i=1}^{n} \|F_{\mathrm{h}m} \times SR_i - F_{\mathrm{h}m} \times HR_i\|_1 \tag{7}$$

where $F_{\mathrm{h}m}$ is the background removal mask.

In this paper, the content loss in the reconstruction loss function is indeed calculated by utilizing the VGG16 network [13]. This process extracts the high-level semantic features of the image. Following this, the L_1 paradigm constraints are applied to these high-level semantic feature maps as per the methodology outlined in Eq. (8).

$$L_{perce} = \frac{1}{N}\sum_{i=1}^{N} \|VGG(SR_i) - VGG\left(HR_i\right)\|_1 \tag{8}$$

To enhance the perceptual realism of the reconstructed image, we incorporate adversarial loss into the loss function, as described in Eqs. (9) and (10).

$$L_{gen} = \frac{1}{N}\sum_{i=1}^{N} \log\left(1 - D\left(G\left(LR_i\right)\right)\right) \tag{9}$$

$$L_{dis} = \frac{1}{N}\sum_{i=1}^{N} \left[\log D\left(HR_i\right) + \log\left(1 - D\left(G\left(LR_i\right)\right)\right)\right] \tag{10}$$

where L_{gen} and L_{dis} are utilized to alternately update the parameters of the generator in Fig. 2 and the discriminator.

The overall reconstruction loss function L comprises pixel loss, perceptual loss, and adversarial loss, as described in Eq. (11).

$$L = L_{pixel} + 0.006L_{perce} + 0.001L_{gen} \tag{11}$$

3. Experiments

3.1 Datasets

The training and testing phases of this study utilized three facial datasets, namely CelebA [8], Helen [9], and FFHQ [10]. CelebA dataset was divided into three subsets: a training set consisting of 162,770 sets of LR/HR face images, a validation set with 19,867 sets and a test set with 19,962 sets. The FFHQ dataset was divided into a training set containing 60,000 sets of LR/HR face images and a test set comprising 10,000 sets of LR/HR face images. Lastly, the Helen dataset was split into a training set consisting of 2,000 sets of LR/HR face images and a test set with 330 sets of LR/HR face images.

3.2 Experimental Setup

The method presented in this study leverages the PyTorch framework for implementation. The algorithms were trained and validated using an NVIDIA TITAN Xp, and all the SRResNet, BiSeNet, and our self-enhancing module were trained simultaneously. The batch size for the training data is set at 16. The resolution of the HR face images from CelebA and Helen datasets is set at 128×128, with a downsampling factor of 8. The optimization process incorporates the Stochastic Gradient Descent (SGD) algorithm, utilizing the Adam parameter for learning rate (learning-rate=0.001),

Table 1 Quantitative results on CelebA and Helen datasets.

algorithms	CelebA		Helen	
	PSNR	SSIM	PSNR	SSIM
Bicubic	23.75	0.624	23.16	0.633
PULSE	25.96	0.742	25.45	0.734
SPARNet	26.43	0.764	26.14	0.756
DICGAN	26.72	0.793	26.24	0.784
GFPGAN	26.68	0.763	26.07	0.754
Ours	26.76	0.806	26.26	0.780

Fig. 5 Comparison of SR effects between the proposed algorithm and other SR algorithms on CelebA dataset.

and the weight decay coefficient (betas = (0.5,0.0.9)).

3.3 Comparative Experiments

In this subsection, we present both quantitative and qualitative evaluations of the algorithms proposed in our study, specifically focusing on a magnification factor of 8.

The quantitative evaluations involve comparing the Peak Signal-to-Noise Ratio (PSNR) and Structure Similarity Index Measure (SSIM) of our proposed algorithms with existing methods on the CelebA and Helen datasets. For the qualitative evaluations, we show the SR face images reconstructed by various algorithms on the CelebA and Helen test sets.

The existing methods used for comparison in the quantitative experiments include PULSE [3], SPARNet [1], DICGAN [4], and GFPGAN [5]. The results of these experiments are presented in Table 1. According to the results presented in Table 1, our proposed FSR algorithm achieves the highest PSNR values on the CelebA and Helen datasets. It outperforms DICGAN by 0.04 dB and 0.02 dB, respectively. Moreover, our algorithm also achieves the best SSIM results on the CelebA dataset, surpassing DICGAN by 0.013 dB.

In the qualitative experiments, we evaluated the performance of various algorithms, including SPARNet, DICGAN, GFPGAN, and PULSE. It is worth noting that the PULSE produces the obvious wrong identity-aware details, as can be observed in Fig. 5.

In comparison, our method aims to faithfully retain the identity information from the original image while also recovering some of the finer details. As a result, the reconstruction of the facial area is superior, which helps mitigate the presence of artifacts in the image. This indicates that our algorithm is successful in preserving the identity information of the face while enhancing the overall quality of the image.

Fig. 6 Super-resolution effect of different models.

3.4 Ablation Experiments

The dual-branch enhancement structure was removed from the network to assess the effectiveness of this particular module. Similarly, the structure extraction module was eliminated from the hair branch to evaluate its impact. Ablation experiments were conducted using the FFHQ dataset, where the LR face image resolution was set at 32×32 and the HR face image resolution was set at 256×256.

Figure 6 illustrates the results of these ablation experiments. The images shown are:

(a) The Interpolated Low-Resolution (ILR) face image obtained after magnifying the LR image 8 times using bicubic interpolation.
(b) The SR face image reconstructed by SRResNet only.
(c) The SR face image reconstructed by the network after removing the structure extraction module from the hair branch.
(d) The SR face image reconstructed by the network using the method proposed in this study.
(e) The HR face image, which serves as the ground truth.

By comparing these images, we can assess the impact of removing specific modules from the network and determine the effectiveness of the proposed method in reconstructing the HR face image.

Compared to (b), both (c) and (d) exhibit superior recovery of the facial region, effectively demonstrating the enhancement of the facial portion of the image due to the dual-branch structure. When compared to (c), the hair region in (d) is rich in texture, featuring distinct hair edges and an overall glossy, bright color. This effectively underlines the significance of the structural extraction module in extracting hair directionality and texture.

4. Conclusion

In this paper, we proposed a semantic-based two-branch

FSR algorithm. Firstly, the resolution of the LR face image was enhanced using the residual super-resolution module. Then, the semantics of the pixels in the enhanced face image were identified by using the segmentation network module, which resulted in a hierarchical image. Then, the hierarchical image was fed into the proposed dual-branch structure network. Finally, we design a novel structure extraction module and a self-enhancing module to further enhance the high frequency detail information and improve the overall reconstruction quality. Extensive experiments are implemented on face datasets, which fully demonstrates that our proposed method can match and exceed the state-of-the-art performance in both quantitative and qualitative measurements.

References

[1] C. Chen, D. Gong, H. Wang, Z. Li, and K.-Y.K. Wong, "Learning spatial attention for face super-resolution," IEEE Trans. Image Process., vol.30, pp.1219–1231, 2021.
[2] I. Goodfellow, J. Pouget-Abadie, M. Mirza, et al., "Generative adversarial nets," Advances in Neural Information Processing Systems, 27, 2014.
[3] S. Menon, A. Damian, S. Hu, N. Ravi, and C. Rudin, "Pulse: Self-supervised photo upsampling via latent space exploration of generative models," Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp.2437–2445, 2020.
[4] C. Ma, Z. Jiang, Y. Rao, J. Lu, and J. Zhou, "Deep face super-resolution with iterative collaboration between attentive recovery and landmark estimation," Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp.5569–5578, 2020.
[5] X. Wang, Y. Li, H. Zhang, and Y. Shan, "Towards real-world blind face restoration with generative facial prior," Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp.9168–9178, 2021.
[6] C. Ledig, L. Theis, F. Huszár, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, Z. Wang, and W. Shi, "Photo-realistic single image super resolution using a generative adversarial network," Proc. IEEE Conference on Computer Vision and Pattern Recognition, pp.4681–4690, 2017.
[7] C. Yu, J. Wang, C. Peng, C. Gao, G. Yu, and N. Sang, "BiSeNet: Bilateral segmentation network for real-time semantic segmentation," Proc. European Conference on Computer Vision, pp.325–341, 2018.
[8] Z. Liu, P. Luo, X. Wang, and X. Tang, "Deep learning face attributes in the wild," Proc. IEEE International Conference on Computer Vision, pp.3730–3738, 2015.
[9] V. Le, J. Brandt, L. Zhe, L.D. Bourdev, and T.S. Huang, "Interactive facial feature localization," Proc. European Conference on Computer Vision, pp.679–692, 2012.
[10] T. Karras, S. Laine, and T. Aila, "A style-based generator architecture for generative adversarial networks," Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp.4396–4405, 2019.
[11] C. Liu, "Gabor-based kernel PCA with fractional power polynomial models for face recognition," IEEE Trans. Pattern Anal. Mach. Intell., vol.26, no.5, pp.572–581, 2004.
[12] X. Wang, R. Girshick, A. Gupta, and K. He, "Non-local neural networks," Proc. IEEE Conference on Computer Vision and Pattern Recognition, pp.7794–7803, 2018.
[13] Q. Cao, L. Shen, W. Xie, O.M. Parkhi, and A. Zisserman, "VGGFace2: A dataset for recognising faces across pose and age," 2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018), pp.67–74, 2018.