IEEE | Transactions on Intelligent Transportation Systems

Joint Edge Information and Attention Aggregation Network for Face Super-Resolution

Journal:	*Transactions on Intelligent Transportation Systems*
Manuscript ID	T-ITS-22-03-0713
Manuscript Type:	Special issue on Internet of Things in Intelligent Transportation Infrastructure
Date Submitted by the Author:	17-Mar-2022
Complete List of Authors:	Gao, Guangwei; Nanjing University of Posts and Telecommunications, Institute of Advanced Technology Tang, Lei; Nanjing University of Posts and Telecommunications, College of Automation Wu, Fei; Nanjing University of Posts and Telecommunications, School of Transportation Chang, Houyou; Nanjing Xiaozhuang College, Key Laboratory of Trusted Cloud Computing and Big Data Analysis Guest Editor - SI (Responsible Artif Intell) Lu, Huimin; Kyushu Institute of Technology, Electrical Engineering Yu, Yi; National Institute of Informatics, Digital Content and Media Sciences Research Division
Keywords:	Image processing
Abstract:	Face super-resolution (SR) is a field-specific image super-resolution problem that is often desired in various industrial application scenes such as video surveillance and identification systems. Most deep learning steered solutions designed for face images often take advantage of the facial priors (i.e., face parsing and landmark) to restore elaborated facial components and have achieved promising performance. However, these methods usually require abundant extra manually labeled data and long training time. In this work, we introduce a fresh Joint Edge Information and Attention Aggregation Network for the face SR problem, dubbed as JEANet, established on our carefully designed Face Attention Aggregation Modules (FAAMs), which is composed of the parallel connection of both channel-wise and spatial-wise features. Specifically, we incorporate an attention fusion mechanism to the residual blocks and meanwhile use edge blocks to extract edge information. Moreover, we interpolate adaptive shortcuts to the reconstruction parts at multiple scales. This structure urges the convolutional operations to flexibility distill information related to the primary facial structures and progressively supply edge information to assemble local and global useful structures. Benefiting from the adaptive shortcuts, our JEANet can restore detailed textures of images from shallow features in an optimal way. Qualitative and quantitative evaluations have shown that the designed method has the superiority of recovering high realistic face images over some competitive face SR methods.

Joint Edge Information and Attention Aggregation Network for Face Super-Resolution

Guangwei Gao, *Member, IEEE,* Lei Tang, Fei Wu, Heyou Chang Huimin Lu, *Senior Member, IEEE,* and Yi Yu,*, *Senior Member, IEEE*

***Abstract*—Face super-resolution (SR) is a field-specific image super-resolution problem that is often desired in various industrial application scenes such as video surveillance and identification systems. Most deep learning steered solutions designed for face images often take advantage of the facial priors (i.e., face parsing and landmark) to restore elaborated facial components and have achieved promising performance. However, these methods usually require abundant extra manually labeled data and long training time. In this work, we introduce a fresh Joint Edge Information and Attention Aggregation Network for the face SR problem, dubbed as JEANet, established on our carefully designed Face Attention Aggregation Modules (FAAMs), which is composed of the parallel connection of both channel-wise and spatial-wise features. Specifically, we incorporate an attention fusion mechanism to the residual blocks and meanwhile use edge blocks to extract edge information. Moreover, we interpolate adaptive shortcuts to the reconstruction parts at multiple scales. This structure urges the convolutional operations to flexibility distill information related to the primary facial structures and progressively supply edge information to assemble local and global useful structures. Benefiting from the adaptive shortcuts, our JEANet can restore detailed textures of images from shallow features in an optimal way. Qualitative and quantitative evaluations have shown that the designed method has the superiority of recovering high realistic face images over some competitive face SR methods.**

***Index Terms*—Attention mechanism, Edge block, Residual block, Face super-resolution**

I. Introduction

The target of general image super-resolution is to yield high-resolution (HR) realistic images from the corresponding low-resolution (LR) inputs suffering from different degradation [1]–[4]. This technology has been widely used in many industrial scenes [5]–[7]. Face super-resolution (SR), meanwhile dubbed as face hallucination, has drew much attention in visual processing community, and vast approaches have been presented in past several years [8]–[15]. On top of the above, face SR also needs to focus on the recovery of the key face components and facial details. Previous face SR works typically exploit face-specific priors, such as facial landmarks [16], parsing maps [16], [17], and facial heatmaps [18], and have shown that those geometry facial priors are pivotal to recover accurate facial shape and details. Joint learning with those additional priors indeed assists enhancing the performance of the face SR task. However, there are manly two drawbacks need to be considered, namely (1) labeling the data requires extra effort, and (2) forecasting face priors from degraded observations is explicitly a tough task.

Given the faithful restoration of the facial details from the degraded input is crucial for many applications, such as face recognition and identification. Thus, we should refine other useful facial information to take place of the above sophisticated facial priors. Recently, face SR methods [19], [20] exploited edge information to alleviate the latent distortion phenomenons in SR problem. Such tendencies verify that the adoption of edge information is an alternative scheme to improve the quality of the recovered images. Particularly, these solutions can be breaked down into two processes: edge distillation and SR conversion. First, the edge extraction network yields the edge features from input LR face and coarse super-resolved ones. Second, the attained edge map is transmitted to the following SR procedure to compensate the desired high-frequency information, resulting in an increase in the SR performance. Different from previous methods [8], [9], [16], [21], which need additional networks to obtain the facial priors, the edge detection network is composed of an individual convolutional layer and an individual average-pooling layer, and have exhibited higher performance in providing structural information. Meanwhile, in previous face SR tasks, the hourglass block is proven to be able to capture informative information at multiple scales [22] and has achieved prominent progress in face image analysis problems, such as face parsing and face alignment [23].

In this work, we carefully design the Face Attention Aggregation Module (FAAM) to establish the Joint Edge Information and Attention Aggregation Network (JEANet) for face SR task. In addition, we introduce edge information into our network for elaborately reconstructing facial components (one example is given in Fig. 1). To effectively capture multi-scale information, we extract informative features closed to the face structures using an attention branch after an hourglass block. We also associate channel-wise features and spatial-wise features through parallel connection to calculate complementary attention. For the idea of achieving the channel attention map, we employ the solution designed by Woo et al. [24] which adopts both max-pooled and average-pooled

*Corresponding author.
G. Gao is with the Institute of Advanced Technology, Nanjing University of Posts and Telecommunications, China (e-mail: csggao@gmail.com).
L. Tang and F. Wu are with the College of Automation, Nanjing University of Posts and Telecommunications, China (e-mail: tl_njupt@163.com, wufei_8888@126.com).
H. Chang is with Key Laboratory of Trusted Cloud Computing and Big Data Analysis, Nanjing XiaoZhuang University, China (e-mail: cv_hychang@126.com).
H. Lu is with the Department of Mechanical and Control Engineering, Kyushu Institute of Technology, Japan (e-mail: dr.huimin.lu@ieee.org).
Y. Yu is with the Digital Content and Media Sciences Research Division, National Institute of Informatics, Japan (e-mail: yiyu@nii.ac.jp).

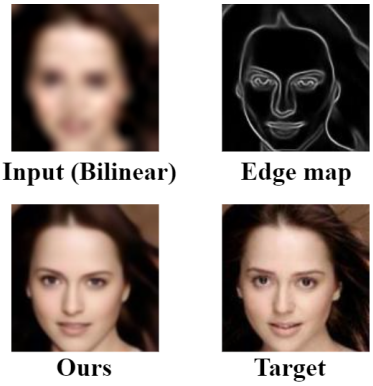

Fig. 1: SR result by our proposed JEANet with scale factor 8.

features simultaneously. Spatial attention features in various FAAMs of the network can be proven to concentrate on diverse types of face structure. For instance, attention features in shallower layers mainly focus more on explicit textures such as hairs, while those in deeper layers mainly concentrate more on contour structures such as mouth and eyes. To reduce the affect caused by the gradient divergence, we further incorporate adaptive shortcuts which add adaptive weights to realize trade-off between quality and module parameters at multiple scales to mitigate the degradation of model. Our network has bigger weights in the shallow layers, especially in the shortcut branch, totally different from the widely used residual branch. It is worthy that more significant information in the shallow layers can be transmitted to the deeper ones of the network. The primary contributions of this work can be enumerated as

- For better feature extraction, we design the Face Attention Aggregation Module (FAAM), which consists of the parallel connection of channel-wise and spatial-wise features. FAAM can significantly distill the vital face properties (i.e., face components and face outlines) and elaborately enhance the results of face SR.
- We interpolate adaptive shortcuts into our framework to add adaptive weights to branches of the reconstruction part at multiple scales. Experiments demonstrate that by considering the adaptive shortcuts, our network results in better performance.
- We propose an efficient Joint Edge Information and Attention Aggregation Network (JEANet) for face SR. We introduce edge information instead of complicated supervisions (e.g. face landmarks and face parsing maps), which enables our network to elaborately obtain clear face images and achieve better SR performance.

II. Related Works

A. Face Image Super-Resolution

Lately, deep learning steered face SR solutions have derived prominent progress in a variety of image quality enhancement tasks. Yu et al. [25] incorporated a flexible deep discriminative network which can obtain desired faces from very LR input face counterparts. Then Huang et al. [26] resorted to the wavelet field and presented a framework that forecasts wavelet coefficients based on the HR images. Later, Yu et al. [27] further implanted informative features in the procedure of face SR. Some advanced face SR solutions also category the methods into local and global ones. Tuzel et al. [28] designed a framework that consists of two parts: one super-resolves faces based on the global constraints, while the other focuses on the local details enhancement. Cao et al. [29] proposed to introduce reinforcement learning scheme to stress attended areas and explore a local enhancement network for detail reconstruction. Because of the fact that face SR is a field-specific image SR mission, facial structural priors are leveraged in many classic image SR solutions. Yu et al. [18] concatenated deep features with facial component heatmaps in the intermediate part of the whole framework. Chen et al. [16] desired to embed facial parsing maps and facial landmark heatmaps simultaneously. Kim et al. [8] designed a facial landmark heatmaps steered attention loss and then utilized it to guide the training of a progressive generator. Although additional supervisions can help to strengthen the SR performance, the network requires extra labels (e.g., face parsing maps and landmarks) to guide the training of the model. Also, it is fussy to get the explicit labels in real applications. Moreover, the lack of ability to assign weights to residual branches may also lead to severe distortions.

B. Attention Mechanism

Recently, attention weighting scheme has been broadly utilized in many practical vision application fields, such as image captioning [30], [31], image classification [32], [33], and visual question answering [34], [35]. The primary issue is to adjust features (maps) through a weight matrix to enhance informative maps while restrain less important ones [35]. He et al. [33] designed the so-called Squeeze-and-Excitation architecture, which exploits a channel-wise weighting scheme and demonstrated remarkable manifestation enhancements. Wang et al. [36] presented a trunk-and-mask weighting concept to the residual network for general image classification task. Recently, attention weighting scheme has also been exploited in image generation problems. Woo et al. [24] presented a powerful convolutional block attention module (CBAM), that subsequently obtains weighting maps along the spatial and channel dimensions simultaneously. They employed pooling to extract spatial-wise information which may lead to the loss of useful low- and middle-level information. In comparison, our spatial attention scheme is explicitly presented to take full advantage of the multi-scale features. Furthermore, unlike the serial connection in CBAM, the channel attention and the spatial attention scheme are integrated parallelly to enable the model benefit from the contextual information.

C. Edge Features in Super-Resolution

Generally, a natural image could be categorized into two parts: low-frequency and corresponding high-frequency components. The latent high-frequency part contains informative features, such as edges, to recover desired structures of the observations. Inspired by the edge information, some competitive methods [19], [37], [38] have concentrated on the recovery of the high-frequency part. The extracted edge maps were

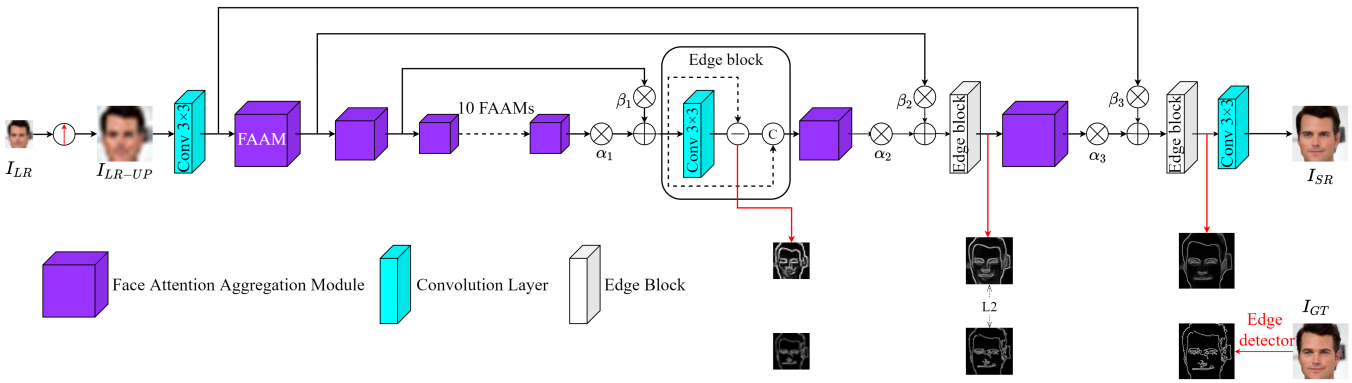

Fig. 2: Network architecture of our proposed JEANet.

usually utilized to promote the quality of the SR images in [37]. The definition of SR formula in [37] was given as an image inpainting task. Fang et al. [19] also used a soft-edge assisted framework to afford edge maps to the SR network. In our face SR network, we insert edge blocks to provide edge assisted information to progressively aggregate local and global structural property for better reconstruction.

III. Proposed Method

A. Overview

Our designed network architecture is comprised of sixteen FAAMs (see 3.2 for details) and other two independent convolutional layers located in the front and back location of the proposed network. The first three FAAMs are used for downscale operations, the middle ten FAAMs are used for feature extraction, and the last three FAAMs constitute the upscale parts. Considering that upscale operations are intuitively proven to be more susceptive to the attention and extracting information only by pooling operations may lead to loss of the shape and edge detail, three edge blocks (see 3.3 for details) are embedded after each FAAM in upscale parts. The overall network framework of our JEANet is depicted in Fig. 2. Let's denote I_{LR}, I_{SR} and I_{HR} as the LR input face image, the output SR image, and the ground truth HR image, respectively. We first upsample I_{LR} to the same size as I_{HR} through bicubic interpolation and denote it as I_{LR-UP}. Then we feed I_{LR-UP} to JEANet to generate I_{SR}. Besides, we incorporate adaptive shortcuts to add adaptive weights at multiple scales to achieve the trade-off between the reconstruction quality and module parameters. As the network is gradually optimized, these multiplicative factors (i.e., $\alpha_k, \beta_{k,k\in(1,2,3)}$) as shown in Fig. 2) will alter adaptively to obtain the optimal performance.

B. Face Attention Aggregation Module

Considering that some face components (e.g., eyes, eyebrows, nose, and mouth) may play a key role in face SR problem, we design an attention aggregation mechanism which exploits both spatial-wise and channel-wise attention to allow our method concentrate more on the informative properties. First of all, the task need to be completed is how to generate the attention maps and aggregate them with the convolutional layers. When it comes to the face SR, we often focus on the high-level information of the face that promotes the network to learn how face looks, and ignore the low-level one. If low-level feature is absent in the network, the ability of the network to learn local details will be worse. Hence, it is essential for the attention aggregation scheme to be capable of learning from multi-scale aspects. Furthermore, residual compositive modules have achieved significant progress in both ecumenical image SR task [39], [40] and specific face SR task [8], [9]. Integrating attention mechanism with residual blocks is profitable to promote the performance of the network.

On account of the above considerations, we present a Face Attention Aggregation Module (FAAM) (see Fig. 3), which enriches original residual blocks by incorporating attention branches. In view of hourglass block's ability to extract multi-scale features and excellent performance in face tasks, we adopt it before the attention branches to facilitate the generation of the attention maps. The attention branches can be divided into a channel branch and a spatial branch where the channel branch and spatial branch interact in parallel to further distill and enhance face information. Feature branch contains convolutional layers, batch normalization and PReLU [41] where upscaling and downscaling can be completed by adding a nearest-neighbour upsampling layer and by changing the step size of the convolutional layer to 2, respectively. Finally, these FAAMs are stacked together to acquire richer feature information and improve the learning ability of the network. Denote by x_{j-1} as the feature input of the j-th FAAM, the channel attention map μ_{j} and spatial attention map ν_{j} can be defined as

$$f_j = F_{feat}(x_{j-1}), \tag{1}$$

$$\mu_j = \sigma(F_{attc}(f_j)), \tag{2}$$

$$\nu_j = \sigma(F_{atts}(f_j)), \tag{3}$$

where f_j denotes the output of the so-called feature branch F_{feat}, σ is the sigmoid function, F_{attc} and F_{atts} are the channel branch and spatial branch respectively. Then the fusion operation can be defined as

$$f_t = [f_j \otimes \mu_j, f_j \otimes \nu_j], \tag{4}$$

where f_t is the fusion result of the features and attention maps, $\otimes$ and [,] denote the procedure of element-wise multiplication

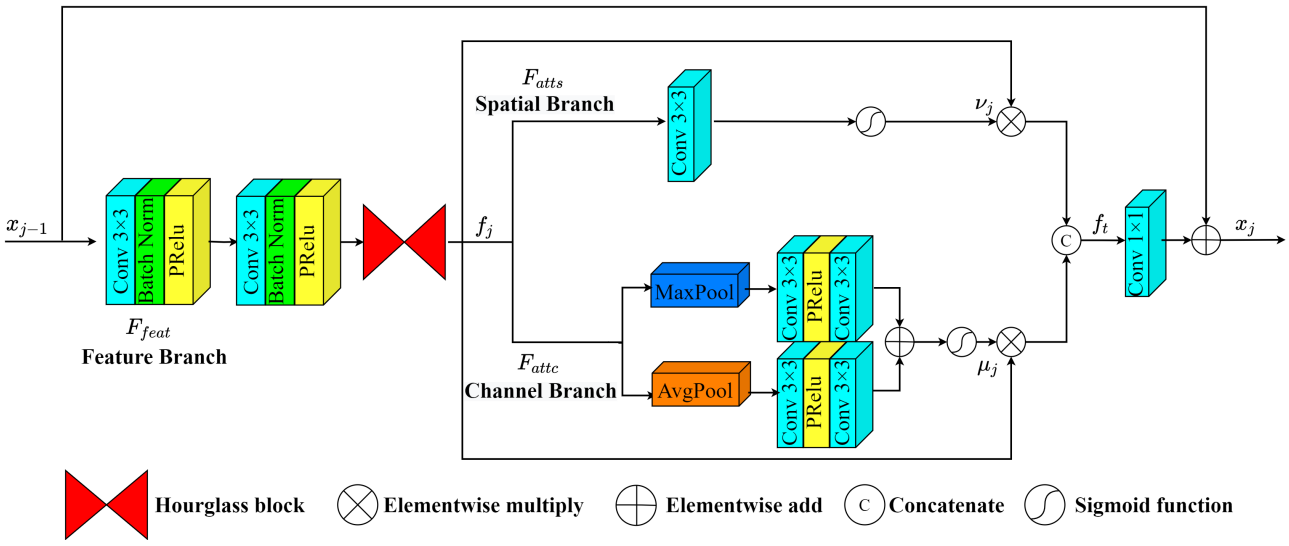

Fig. 3: Face Attention Aggregation Module.

and concatenation respectively. Ultimately, the export of the j-th FAAM is formulated as

$$x_j = x_{j-1} + Conv1(f_t), \tag{5}$$

here $Conv1$ indicates a convolutional layer with the kernel size of 1×1 whose role is to reduce the number of the feature maps.

C. Edge Block

Edge information has been adopted to promote image qualities in multiple visual applications problems, e.g., single image SR, image inpainting, and image denoising. Undoubtedly, edge property is a significant element which not only exquisitely obtains facial details from LR faces but also deal with the inherent issue of L_2 loss that desires to yield smoothed images in SR problem. Furthermore, the edge property can compensate for lost information in high-frequency parts, such as shapes and edges, to improve the qualities of the SR images. To fully distill and adopt the edge property, we adopt the edge block proposed by Kim et al. [42]. The edge block is generally composed by a simplex convolutional layer, a simplex average pooling layer and PRelu layer as depicted in Fig. 4. Firstly, it performs average-pooling operation like a low-pass filter in each embedding scale to attain the smoothed features. Then, the edge block yields high-frequency parts by subtracting the smoothed features from the inchoative ones and further decreases the account of the subtracted maps by the operation of $Conv1$. Ultimately, the subtracted edge maps are combined with the acquired features to transmit the edge property to the following FAAM.

D. Loss Functions

The loss functions in the training of our JEANet consists of 1) pixel loss that constraints the outputs I_{SR} approximate the ground truth I_{HR} as much as possible and 2) edge loss for reconstructing the facial details.

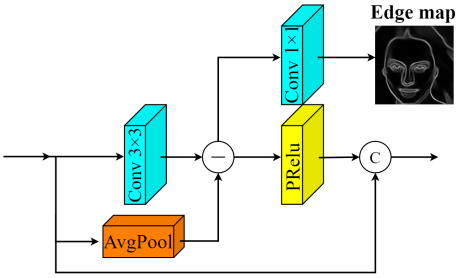

Fig. 4: Edge Block.

1) **Pixel loss.** Actually, L_2 loss gives a weak constraint for small errors while a strong constraint for large errors, overlooking the effect of the content itself. Moreover, the convergence property of the L_2 loss is elaborately worse than that of the L_1 loss. Therefore, the loss function in this work can be formulated as

$$I_{pixel} = \| I_{HR} - I_{SR} \|_1, \tag{6}$$

here $\| \cdot \|_1$ represents the L_1 norm, I_{SR} and I_{HR} are the final output high-quality face image through JEANet and the original HR referenced image respectively.

2) **Edge loss.** Deep learning strategies have achieved prominent improvements in SR tasks. Nevertheless, many of these schemes fail to recover faithful structures as they are usually over-smoothed or blurry. Hence, adding the edge information intuitively compensates for the missing high-frequency parts, such as shapes and edges, to improve the image qualities. To update the edge block, we generate the edge map of HR images using the canny edge detector [39] as the ground truth. The edge property steered loss function is given as

$$L_{edge} = \frac{1}{r} \| C(I_{HR}) - E(I_{LR}) \|_2, \tag{7}$$

where $\|\cdot\|_2$ denotes the L_2 norm, C denotes the canny edge detector, and E represents the edge block. Throughout the network, we consider to use edge information in multi-scales with the scaling factor r to be set as 1, 2, and 4.

Finally, the whole loss function is elaborately defined as

$$L_{total} = \lambda_1 L_{pixel} + \lambda_2 L_{edge}, \tag{8}$$

where λ_1 and λ_2 are the weighting factors for pixel loss and edge loss respectively. We set $\lambda_1 = 1$ and $\lambda_2 = 0.01$ empirically.

IV. Experimental Evaluations

A. Datasets and Metrics

We conduct evaluations on CelebA [43], which is a large-scale face dataset containing about 0.2 million face samples. Inspired by the same previous protocols, we select 162,770 faces to build the training set, 19,867 faces to build the validation set, while 19,962 faces to form the testing set. To further verify the expansibility of our JEANet, we transmit the pre-trained model on CelebA to conduct evaluation on the testing set of Helen [44]. The performance index between the recovered faces and the ground truth ones are given by SSIM and PSNR [45], which are performed on the preferential Y channel in the switched YCbCr feature space.

B. Implementation Guides

To get the HR face images from CelebA dataset, we firstly search faces using the MTCNN [46] tool, and then extract the rough face areas without any pre-alignment operations. Then, we choose face images have the size larger than 128×128, and reshape them to the size of 128×128 using the bicubic interpolation operation, and treat them as the HR ground truth. The LR counterparts are yielded through downsampling the HR ones to the size of 16×16. For fear of overfitting, we perform data augmentation scheme using random horizontal flipping, image rotation (90°, 180°, and 270°), and image rescaling (between 1.0 and 1.3). As for the training of the network, we set the batch-size as 16 and use Adam [47] approach with an primal learning rate as 2×10^{-4}. Our implements are conducted based on the Pytorch [48] platform using powerful NVIDIA RTX 3090 GPUs.

C. Ablation Studies

We further conduct some studies to evaluate the effectiveness of the edge block and the adaptive shortcuts. On one hand, to assess the effects of edge information, we remove the edge blocks in every upscaling procedure. On the other hand, we remove the three adaptive shortcuts to evaluate the necessity of the shallow features and adaptive weight assignment. PSNR and SSIM performance CelebA dataset is given in Table I, from which we can observe that when the JEANet loses the guidance of edge information, SR quality decreases due to the weakness of the ability to capture high-frequency components. A faint enhancement can also be observed since the interpolation of adaptive shortcuts. The reason is that transferring shallow features directly to the reconstruction part preserve low-frequency information. In addition to benefit from the adaptive assignment of weights, the network can filter out more useful information. Thus, the above results prove the superiority of our proposed JEANet to gain both low-frequency and high-frequency useful information for better reconstruction purpose.

TABLE I: Ablation studies of different modules.

Method	PSNR↑	SSIM↑
JEANet w/o edge blocks	28.15	0.8061
JEANet w/o adaptive shortcuts	28.16	0.8063
JEANet	**28.19**	**0.8079**

TABLE II: Comparison of PSNR and SSIM performance with respective methods on CelebA and Helen. Red/blue indicates the best/second-best results.

Method	CelebA		Helen	
	PSNR↑	SSIM↑	PSNR↑	SSIM↑
SAN [40]	27.43	0.7826	25.46	0.7363
RCAN [1]	27.45	0.7824	25.51	0.7383
HAN [2]	27.47	0.7838	25.40	0.7347
FSRNet [16]	27.05	0.7714	25.45	0.7364
DICNet [9]	—	—	26.01	0.7659
FACN [13]	27.22	0.7802	25.06	0.7189
SPA [10]	27.73	0.7949	26.43	0.7839
JEANet	28.19	0.8079	26.89	0.7989

D. Experimental Evaluations

In this session, we evaluate our JEANet by comparing it with several competitive SR methods on CelebA and Helen datasets. The compared approaches contain three classical general image SR solutions (SAN [40], RCAN [1] and HAN [2]) and four specific face image SR solutions (FSRNet [16], DICNet [9], FACN [13] and SPA [10]). To make a fair comparison, we train the models of the compared methods with the same CelebA dataset. The qualitative evaluations are provided in Fig 5 and Fig 6. The results gained by the compared approaches posses distinct shadows on some areas and appear blurry in facial details. In contrast, benefiting from the edge blocks that restore facial structures by progressively concatenating the high-frequency component to the primal feature maps, the super-resolved faces by our proposed JEANet can recover more facial details, especially for the parts of eyes, mouth, and hair. Meanwhile, by integrating attention mechanism and adaptive shortcuts, JEANet is better at capturing feature-rich image regions and obtaining quite better visual effects. The quantitative comparisons are also given in Table II. It could be seen that our JEANet approach obtains the optimal PSNR and SSIM results on two datasets, which further validate the advantage of our method. This indicates that our proposed method performs better in preserving facial structures and obtaining better details than other methods.

V. Conclusions

In this paper, we presented a Joint Edge Information and Attention Aggregation Network (JEANet) for face SR task.

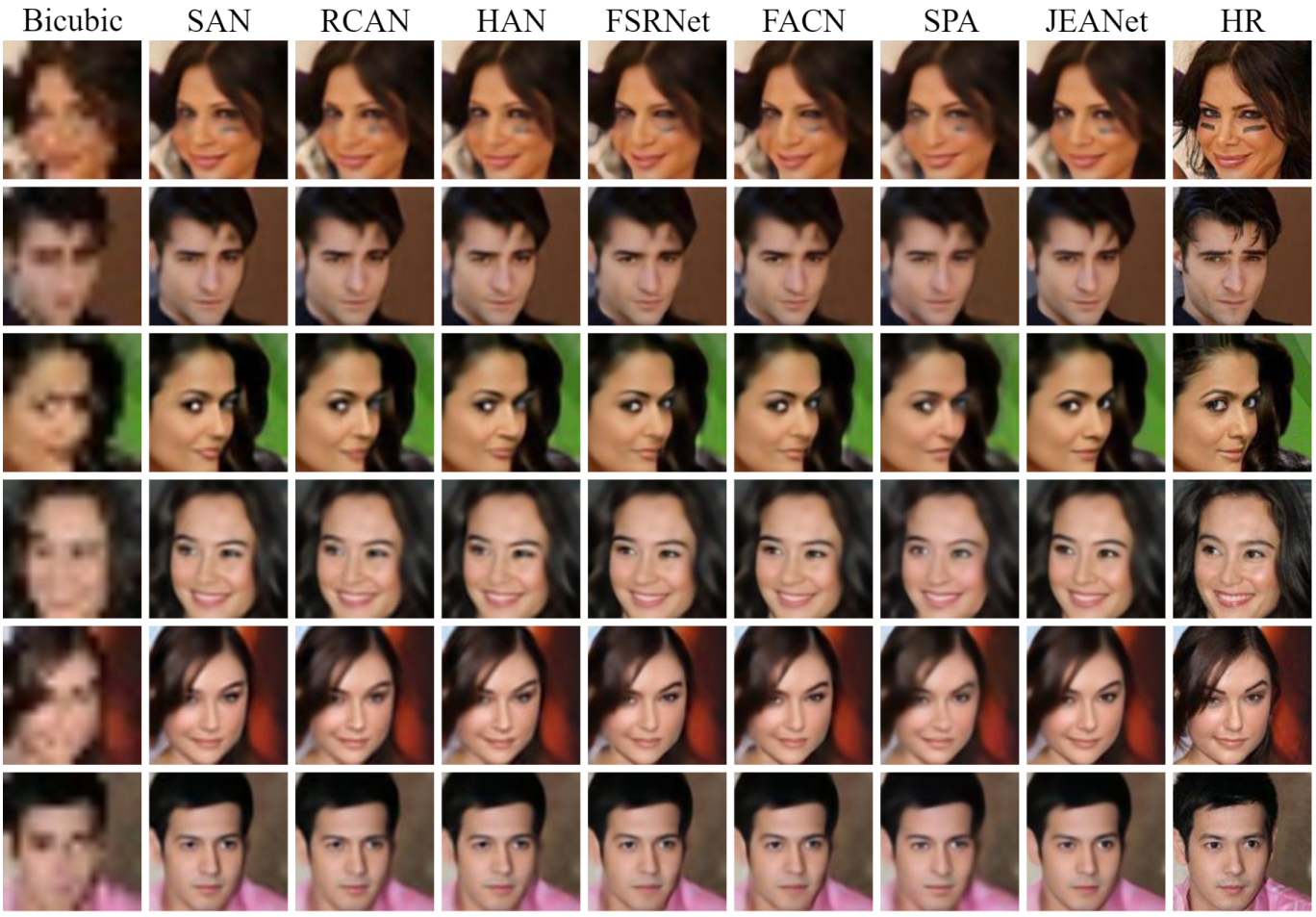

Fig. 5: Visual comparisons on CelebA dataset with scale factor 8. From left to right are successively the SR results of Bicubic, SAN [40], RCAN [1], HAN [2], FSRNet [16], FACN [13], SPA [10], our JEANet and the related ground truth faithful HR references.

In contrast to the conventional solutions which often take advantage of the facial priors (i.e., face parsing and landmark) to restore elaborated facial components, the framework of the proposed JEANet consists of stacking Face Attention Aggregation Models (FAAMs) which utilize the attention fusion mechanism to restrain the network giving more attention to more feature-rich regions. FAAM can distill the vital face properties (i.e., face components and face outlines) and significantly enhance the manifestation of face SR. Moreover, we utilize the adaptive shortcuts to preserve shallow features and progressively provided edge information at multiple scales for mitigating the degradation of the model. Thanks to the above considerations, JEANet is able to recover higher quality face images than various competitive face SR approaches.

REFERENCES

[1] Y. Zhang, K. Li, K. Li, L. Wang, B. Zhong, and Y. Fu, "Image super-resolution using very deep residual channel attention networks," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 286–301.

[2] B. Niu, W. Wen, W. Ren, X. Zhang, L. Yang, S. Wang, K. Zhang, X. Cao, and H. Shen, "Single image super-resolution via a holistic attention network," in *Proceedings of the European Conference on Computer Vision*. Springer, 2020, pp. 191–207.

[3] J. Li, F. Fang, J. Li, K. Mei, and G. Zhang, "Mdcn: Multi-scale dense cross network for image super-resolution," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 31, pp. 2547–2561, 2020.

[4] G. Gao, W. Li, J. Li, F. Wu, H. Lu, and Y. Yu, "Feature distillation interaction weighting network for lightweight image super-resolution," *arXiv preprint arXiv:2112.08655*, 2021.

[5] J. Gu, H. Chen, G. Liu, G. Liang, X. Wang, and J. Zhao, "Super-resolution perception for industrial sensor data," *arXiv preprint arXiv:1809.06687*, 2018.

[6] Z. Ren, H. K.-H. So, and E. Y. Lam, "Fringe pattern improvement and super-resolution using deep learning in digital holography," *IEEE Transactions on Industrial Informatics*, vol. 15, no. 11, pp. 6179–6186, 2019.

[7] S. Ahmadi, G. Thummerer, S. Breitwieser, G. Mayr, J. Lecompagnon, P. Burgholzer, P. Jung, G. Caire, and M. Ziegler, "Multidimensional reconstruction of internal defects in additively manufactured steel using photothermal super resolution combined with virtual wave-based image processing," *IEEE Transactions on Industrial Informatics*, vol. 17, no. 11, pp. 7368–7378, 2021.

[8] D. Kim, M. Kim, G. Kwon, and D.-S. Kim, "Progressive face super-resolution via attention to facial landmark," *arXiv preprint arXiv:1908.08239*, 2019.

[9] C. Ma, Z. Jiang, Y. Rao, J. Lu, and J. Zhou, "Deep face super-resolution with iterative collaboration between attentive recovery and landmark estimation," in *Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition*, 2020, pp. 5569–5578.

[10] C. Chen, D. Gong, H. Wang, Z. Li, and K.-Y. K. Wong, "Learning spatial attention for face super-resolution," *IEEE Transactions on Image Processing*, vol. 30, pp. 1219–1231, 2021.

Fig. 6: Visual comparisons on Halen dataset with scale factor 8. From left to right are successively the SR results of Bicubic, SAN [40], RCAN [1], HAN [2], FSRNet [16], DICNet [9], FACN [13], SPA [10], our JEANet and the related ground truth faithful HR references.

[11] G. Gao, Y. Yu, J. Xie, J. Yang, M. Yang, and J. Zhang, "Constructing multilayer locality-constrained matrix regression framework for noise robust face super-resolution," *Pattern Recognition*, vol. 110, p. 107539, 2021.

[12] G. Gao, Y. Yu, J. Yang, G.-J. Qi, and M. Yang, "Hierarchical deep cnn feature set-based representation learning for robust cross-resolution face recognition," *IEEE Transactions on Circuits and Systems for Video Technology*, 2020.

[13] J. Xin, N. Wang, X. Jiang, J. Li, X. Gao, and Z. Li, "Facial attribute capsules for noise face super resolution," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2020, pp. 12 476–12 483.

[14] X. Hu, W. Ren, J. LaMaster, X. Cao, X. Li, Z. Li, B. Menze, and W. Liu, "Face super-resolution guided by 3d facial priors," in *Proceedings of the European Conference on Computer Vision*, 2020, pp. 763–780.

[15] M. Li, Z. Zhang, J. Yu, and C. W. Chen, "Learning face image super-resolution through facial semantic attribute transformation and self-attentive structure enhancement," *IEEE Transactions on Multimedia*, vol. 23, pp. 468–483, 2021.

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

[17] C. Chen, X. Li, L. Yang, X. Lin, L. Zhang, and K.-Y. K. Wong, "Progressive semantic-aware style transformation for blind face restoration," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 11 896–11 905.

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

[19] F. Fang, J. Li, and T. Zeng, "Soft-edge assisted network for single image super-resolution," *IEEE Transactions on Image Processing*, vol. 29, pp. 4656–4668, 2020.

[20] J. Jiang, Y. Yu, J. Hu, S. Tang, and J. Ma, "Deep cnn denoiser and multi-layer neighbor component embedding for face hallucination," *arXiv preprint arXiv:1806.10726*, 2018.

[21] Y. Yin, J. Robinson, Y. Zhang, and Y. Fu, "Joint super-resolution and alignment of tiny faces," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2020, pp. 12 693–12 700.

[22] A. Mittal, A. K. Moorthy, and A. C. Bovik, "No-reference image quality assessment in the spatial domain," *IEEE Transactions on Image Processing*, vol. 21, no. 12, pp. 4695–4708, 2012.

[23] S. Yang, P. Luo, C.-C. Loy, and X. Tang, "Wider face: A face detection benchmark," in *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 2016, pp. 5525–5533.

[24] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "Cbam: Convolutional block attention module," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 3–19.

[25] X. Yu and F. Porikli, "Ultra-resolving face images by discriminative generative networks," in *Proceedings of the European Conference on Computer Vision (ECCV)*. Springer, 2016, pp. 318–333.

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

[27] X. Yu, B. Fernando, R. Hartley, and F. Porikli, "Super-resolving very low-resolution face images with supplementary attributes," in *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 2018, pp. 908–917.

[28] O. Tuzel, Y. Taguchi, and J. R. Hershey, "Global-local face upsampling network," *arXiv preprint arXiv:1603.07235*, 2016.

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

[30] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel, and Y. Bengio, "Show, attend and tell: Neural image caption generation with visual attention," in *Proceedings of the International Conference on Machine Learning*, 2015, pp. 2048–2057.

[31] L. Chen, H. Zhang, J. Xiao, L. Nie, J. Shao, W. Liu, and T.-S. Chua, "Sca-cnn: Spatial and channel-wise attention in convolutional networks for image captioning," in *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 2017, pp. 5659–5667.

[32] V. Mnih, N. Heess, A. Graves *et al.*, "Recurrent models of visual attention," in *Proceedings of the Advances in Neural Information Processing Systems*, 2014, pp. 2204–2212.

[33] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 2018, pp. 7132–7141.

[34] I. Schwartz, A. G. Schwing, and T. Hazan, "High-order attention models for visual question answering," *arXiv preprint arXiv:1711.04323*, 2017.

[35] D. Yu, J. Fu, T. Mei, and Y. Rui, "Multi-level attention networks for visual question answering," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 4709–4717.

[36] F. Wang, M. Jiang, C. Qian, S. Yang, C. Li, H. Zhang, X. Wang, and X. Tang, "Residual attention network for image classification," in *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 2017, pp. 3156–3164.

[37] K. Nazeri, H. Thasarathan, and M. Ebrahimi, "Edge-informed single image super-resolution," in *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, 2019, pp. 0–0.

[38] K. Kim and S. Y. Chun, "Sredgenet: Edge enhanced single image super resolution using dense edge detection network and feature merge network," *arXiv preprint arXiv:1812.07174*, 2018.

[39] J. Canny, "A computational approach to edge detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 8, no. 6, pp. 679–698, 1986.

[40] T. Dai, J. Cai, Y. Zhang, S.-T. Xia, and L. Zhang, "Second-order attention network for single image super-resolution," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 11 065–11 074.

[41] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on imagenet classification," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 1026–1034.

[42] J. Kim, G. Li, I. Yun, C. Jung, and J. Kim, "Edge and identity preserving network for face super-resolution," *Neurocomputing*, vol. 446, pp. 11–22, 2021.

[43] Z. Liu, P. Luo, X. Wang, and X. Tang, "Deep learning face attributes in the wild," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 3730–3738.

[44] V. Le, J. Brandt, Z. Lin, L. Bourdev, and T. S. Huang, "Interactive facial feature localization," in *Proceedings of the European Conference on Computer Vision (ECCV)*. Springer, 2012, pp. 679–692.

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

[46] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao, "Joint face detection and alignment using multitask cascaded convolutional networks," *IEEE Signal Processing Letters*, vol. 23, no. 10, pp. 1499–1503, 2016.

[47] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[48] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer, "Automatic differentiation in pytorch," 2017.