Progressive face super-resolution via learning prior information

Yanfei Duan¹, Yintian Liu², Ruixiang Wang, Dengguo Yao and Hang Zhang

College of Software Engineering, Chengdu University of Information Technology, Chengdu, SiChuan,610225, China
¹liuyintian@cuit.edu.cn ²Corresponding author’s e-mail: 865582233@qq.com

Abstract. Face super-resolution is a branch of the field of super-resolution. It is mainly aimed at the reconstruction of face images and distinguishing SR from general images. Face geometry prior information is used to optimize the face super-resolution network, which can generate high-resolution face images with better visual quality from low-resolution face images. In order to further improve the visual quality of reconstructed images, an improved face super-resolution network is proposed. In this paper, the key module of FSRNet is improved and new loss function is added to achieve a better face super-resolution network. Specifically, our job is to:(1) In the generated rough SR face image, we input low-resolution(16 x16) face image, then use the Deconv convolution enlarge images.(2) By introducing heatmap loss, facial attention loss and adversarial training to reduce the artifacts of FSRNet network. (3) We divide the network into two-step training, first train coarse SR network, get SR images quickly, and then we train the rest of the network. The final output is a super-resolution face with high visual quality.

1. Introduction

Facial super-resolution can greatly facilitate face-related tasks. For facial-related applications, such as face parsing [1], face alignment and face recognition [2], it is very important. The single image super-resolution method [3] relies on the loss of pixel-level mean square error (MSE). However, such an approach produces smooth and fuzzy output and lacks texture detail. As a special case of the ordinary image SR, the facial image has a priori knowledge specific to the face, which is not applicable to the ordinary image SR [4]. For example, facial structure information [5] helps to restore facial shape, and facial components [6] can show rich facial details. We have analyzed the role of each module in the FSRNet network and found that the poor visual quality of FSRNET output SR is due to the relatively undeveloped design of the coarse SR network module and the loss function. The FSRNet approach does not adequately consider the facial properties of the area around the landmark. We propose a loss of facial attention, focusing attention on facial details in the landmark area. Reference [7] proposed a compressed version of FAN face alignment network to extract the heat map of key points of the face for supervision and training.

2. Related work

We improved the coarse SR network part of FSRNet to get better facial information quickly. By introducing heatmap loss and facial attention loss to train the network, the network model performance is further improved.
2.1. Improve Coarse SR Network

The input image of FSRNet is the interpolated low-resolution image (128x128x3). We reduced the size of the input image, fed it into the CoarseSR with low-resolution image (16x16x3), and enlarged the image to 128x128x3 in the tail of CoarseSR by use of Deconv. CoarseSR use three residual blocks to extract features. Our modified CoarseSR structure is shown in Figure 1.

![Figure 1. Coarse SR Network.](image1)

2.2. Two step training

We train the CoarseSR firstly and then we train the rest of the network. Different from the FSRNet method, the method in this paper can precisely restore the facial details by imposing constraints on the critical regions of the face. We adopt progressive training method [8-11] and introduce adversarial loss training. We increase the loss of facial attention and restore more accurate facial detail. We used the facial alignment network (FAN) provided by [7] to extract heatmaps to obtain facial attention loss. We calculate the average Peak Signal to Noise Ratio (PSNR) and Structural Similarity (SSIM) to evaluate the proposed method. The improved overall network structure is shown in Figure 3. In the first step, train CoarseSR and learn to enlarge the images by a factor of 4. The 4x images pass through a convolutional layer, are then fed into the corresponding part of the discriminator, and the output is compared with the target image. In the second step, the output image of the first step is fed into the rest of the network. The final SR output is fed into the discriminator and then compared with the target image.

![Figure 2. Network architecture overview.](image2)

3. The objective function

To generate heatmaps suitable for focusing on accurate facial landmarks, we used the distilled FAN [7] network to extract heatmaps of faces for Supervision and training. The introduced distillation FAN network is represented by $F_d$. 
3.1. Facial attention loss

\[ L_{\text{attention}} = \frac{1}{2WH} \sum_{x=1}^{W} \sum_{y=1}^{H} (M^*_{(x,y)} - G(I_{(x,y)}^{LR})) \]  

(1)

Where, \( G \) stands for generator, namely face SR network. In the first step of training, \( G \) represents the coarse SR network. In the second step, \( G \) represents the rest of the network. HR and LR are target face image and input image respectively. Key point attention heat map \( M^* \) is the maximum value of each channel of the target heatmap \( M \) generated from the target face image. Heatmap \( M \) is normalized \([0,1]\). The size of heatmap \( M \) is \( N \times W \times H \), where \( N \) is the number of key points. \( W \) and \( H \) are the width and height of the image. Without special instructions, both \( W \) and \( H \) are 128 pixels. To focus attention on images with sufficient information, we added facial attention loss to both steps.

3.2. MSE Loss

We use the per-pixel mean square error (MSE) loss to minimize the distance between the target image HR and the super-resolution image SR.

\[ L_{\text{pixel}} = \frac{1}{2WH} \sum_{x=1}^{W} \sum_{y=1}^{H} (I_{(x,y)}^{HR} - G(I_{(x,y)}^{LR}))^2 \]  

(2)

3.3. Prior loss

Prior estimation network \( P \) is responsible for predicting face prior information. By minimizing the distance between the generated face prior information and the target face prior information provided by the data set, it provides the SR network with prior information specific to face structure, and its loss is defined as follows:

\[ L_{\text{prior}} = \frac{1}{2WH} \sum_{x=1}^{W} \sum_{y=1}^{H} (I_{(x,y)}^{HR} - P(I_{(x,y)}^{LR}))^2 \]  

(3)

Where, the output of \( P \) network is down sampled, so \( W \) and \( H \) are 64 pixels. \( I_{(x,y)}^{HR} \) is the image output after passing through the coarse SR network.

3.4. Perceived loss

Using the advanced features of the pre-trained VGG16 network to introduce the perceived loss [12], more realistic HR images can be obtained.

\[ L_{\text{perceptual}} = \frac{1}{2WH} \sum_{x=1}^{W} \sum_{y=1}^{H} (\phi(I_{(x,y)}^{HR}) - \phi(G(I_{(x,y)}^{LR})))^2 \]  

(4)

3.5. Adversarial loss

It is well known that generative adversarial network (GAN) shows great strength in super-resolution and has been proved to be effective in restoring high-fidelity and realistic visual images. We established a discriminator \( D \) and used the successful experience of WGAN [13] to stabilize the training process. In WGAN, the loss function is defined as the Wasserstein distance between the target \( f_{\text{IR}} \sim P_t \) distribution and the generated image \( f_{\text{SR}} \sim P_g \) distribution. To further improve the stability of the training, the Gradient Penalty proposed in WGAN-GP [14] was adopted, which forced the Lipschitz-1 condition of the discriminator. \( I \) is the randomly sampled image in the sample from \( P_t \) and \( P_g \). Therefore, the loss function is as follows:

\[ L_{\text{WGAN}} = \mathbb{E}_{I \sim P_t} [D(I^{HR})] - \mathbb{E}_{I \sim P_g} [D(I^{SR})] + \lambda \mathbb{E}_{I \sim P_g} [\| \nabla I D(I^{SR}) \|^2] \]  

(5)
3.6. Heatmap loss
As proposed in reference [15], the structural consistency of facial images is improved by minimizing the distance between the generated image and the heat map of the target image. The heatmap loss function is described as follows:

$$L_{heatmap} = \frac{1}{2NWH} \sum_{w=1}^{n} \sum_{z=1}^{H} \sum_{y=1}^{W} (F_{d}(I^{HR})_{w,z,y} - F_{d}(G(I^{LR})_{w,z,y}))^2$$

(6)

Where $N$ is the number of key points and $F_{d}$ is the pre-trained heat map extraction network.

The training is divided into two steps: the first step is to train the coarse SR network, and the second step is to train the rest of the network. Specific training losses are shown as follows:

$$L_{step1} = \alpha L_{pixel} + \beta L_{perceptual} + \gamma L_{WGAN} + \lambda L_{heatmap} + \eta L_{attention}$$

(7)

$$L_{step2} = \alpha L_{pixel} + \beta L_{perceptual} + \gamma L_{WGAN} + \lambda L_{heatmap} + \eta L_{attention} + \varepsilon L_{prior}$$

(8)

4. Experiment

4.1. Experiment preparation
Data set
In the experiment, we chose the public data set CelebAMask-HQ [16] for the experiment, where 30,000 high-resolution face images were selected from the CelebA data set, and each image had a segmentation mask corresponding to CelebA's face attributes. The face area of the image is cropped and adjusted to 128x128 size as the target image, and bilinear down-sampling is carried out to 16x16 pixels as the LR input. We used 29,000 images from the CelebAMask-HQ dataset for training and the remaining 1,000 images for evaluation.

Training details
We use PyTorch framework to implement our SR network, and Adam optimizer to train the network with the learning rate of 2.5x10-4 and batch size of 16.

4.2. Ablation Study
To verify the validity of the introduced facial attention loss, heatmap loss, and perception loss, we conducted three experiments to conduct ablation studies. PSNR and SSIM were used to assess the impact of each loss. The data in Table 1 are the experimental results. No.1 is the result of only pixel loss and perception loss and adversarial training, No.2 is the result of introducing heatmap loss and No.3 is the result of introducing facial attention loss. The results show that the SR image generated by our method is of higher quality.

| No | Method | PSNR  | SSIM  |
|----|--------|-------|-------|
| 1  | $L_{pixel} + L_{perceptual} + L_{WGAN}$ | 25.36 | 0.725 |
| 2  | $L_{pixel} + L_{perceptual} + L_{WGAN} + L_{heatmap}$ | 25.81 | 0.771 |
| 3  | $L_{pixel} + L_{perceptual} + L_{WGAN} + L_{heatmap} + L_{attention}$ | 26.24 | 0.784 |

To verify the performance of our improved CoarseSR, we compared it with FSRNet's CoarseSR. The low-resolution graph we entered (16x16) reduced the computational complexity by a factor of 16. Table 2 shows the comparison between the improved CoarseSR and the FSRNet's CoarseSR on PSNR and SSIM. Figure 4 shows the results.

| CoarseSR | PSNR | SSIM |
|----------|------|------|
| FSRNet's CoarseSR |     |     |
4.3. Comparison with FSRNet and state-of-the-art

Compare the improved method with some advanced methods. Since our improved method was trained on CelebAMask-HQ, let all methods be tested on CelebAMask-HQ. Figure 4 shows the results of methods, and Table 3 shows the quantitative comparison results on the test set. It can be seen that the improved method in this paper is better than the original FSRNet in all aspects. Visually, The SR image generated by FSRNet contains artifacts and partially blurred facial components. Our method is superior to FSRNet in PSNR and SSIM, and can restore accurate facial features, with a realistic facial visual effect.

Table 3. Comparison of PSNR and SSIM performance with state-of-the-art FSR methods

| Method | PSNR  | SSIM  |
|--------|-------|-------|
| Bicubic| 23.51 | 0.617 |
| RND    | 25.89 | 0.723 |
| VDSR   | 25.93 | 0.729 |
| URDGN  | 25.87 | 0.718 |
| FSRNet | 26.21 | 0.722 |
| Ours   | **26.24** | **0.784** |

5. Conclusions

In this paper, a face super-resolution network is proposed. We use FSRNet's prior estimation network to provide face prior information. We reduce the resolution of the input image, which reduces the computational complexity. Facial attention loss, heatmap loss and adversarial loss training were added to improve network performance. Using the two-step training method, the coarse SR network is first
trained separately to obtain the SR image quickly, and then the rest of the network is trained to generate the fine SR image. Our improved method can generate higher quality face images and perform super-resolution of face images with more accurate face details.

References
[1] Li Y, Liu S, Yang J. (2017) Generative Face Completion. In: Computer Vision and Pattern Recognition. Hawaii. pp. 5892-5900.
[2] Taigman Y, Yang M, Ranzato M. (2014) DeepFace: Closing the Gap to Human-Level Performance in Face Verification. In: Computer Vision and Pattern Recognition. Columbus. pp. 1701-1708.
[3] Dong C, Loy C, He K. (2016) Image Super-Resolution Using Deep Convolutional Networks. IEEE Transactions on Pattern Analysis and Machine Intelligence. 38: 295-307.
[4] Tai Y, Yang J, Liu X. (2017) Image Super-Resolution via Deep Recursive Residual Network. In: Computer Vision and Pattern Recognition. Hawaii. Pp. 2790-2798.
[5] Zhu S, Liu S, Loy C. (2016) Deep Cascaded Bi-Net for Face Hallucination. In: European Conference on Computer Vision. Zürich. pp. 614-630.
[6] Song Y, Zhang J, He S. (2017) Learning to hallucinate face images via component generation and enhancement. In: International Joint Conference on Artificial Intelligence. Melbourne. pp. 4537-4543.
[7] Kim D, Kim M, Won G. (2019) Progressive Face Super-Resolution via Attention to Facial Landmark. In: British Machine Vision Conference. Wales. Pp. 1908-8239.
[8] Ahn N, Kang B, Sohn K. (2018) Image Super-Resolution via Progressive Cascading Residual Network. In: Computer Vision and Pattern Recognition Workshops. Salt Lake City. pp. 904-9048.
[9] Lai W, Huang J, Ahuja N. (2017) Deep Laplacian Pyramid Networks for Fast and Accurate Super-Resolution. In: Computer Vision and Pattern Recognition. Hawaii. pp. 5835-5843.
[10] Karras T, Aila T, Laine S. (2017) Progressive Growing of GANs for Improved Quality, Stability, and Variation. https://arxiv.org/abs/1710.10196.
[11] Wang Y, Perazzi F, Mcwilliams B. (2018) A Fully Progressive Approach to Single-Image Super-Resolution. In: Computer Vision and Pattern Recognition Workshops. Salt Lake City. pp. 977-97709.
[12] Johnson J, Alahi A, Li F. (2016) Perceptual Losses for Real-Time Style Transfer and Super-Resolution. In: European Conference on Computer Vision. Zürich. pp. 694-711.
[13] Arjovsky M, Chintala S, Bottou L. (2017) Wasserstein Generative Adversarial Networks. In: International Conference on Machine Learning. Sydney. 214-223.
[14] Wei X, Gong B, Liu Z. (2018) Improving the Improved Training of Wasserstein GANs: A Consistency Term and Its Dual Effect. https://arxiv.org/abs/1803.01541.
[15] Bulat A, Tzimiropoulos G. (2018) Super-FAN: Integrated Facial Landmark Localization and Super-Resolution of Real-World Low Resolution Faces in Arbitrary Poses with GANs. In: Computer Vision and Pattern Recognition. Salt Lake City. pp. 109-117.
[16] Lee C, Liu Z, Wu L. (2019) MaskGAN: Towards Diverse and Interactive Facial Image Manipulation. https://arxiv.org/abs/1907.11922.