Deep Learning Face Super-Resolution with Facial Prior
VE581 Final Project

Wang Xining[*]
UM-SJTU JI
Shanghai
wxn.sjtu@gmail.com

Hu Chong[*]
UM-SJTU JI
Shanghai
jackchonghu@gmail.com

Abstract

Face Super-Resolution (SR), a.k.a. face hallucination, aims to generate a High-Resolution (HR) face image from Low-Resolution (LR) input. SR can be cooperated with other face-related tasks, such as face recognition, face alignment and face parse. Both traditional statistical analysis and image processing algorithm and deep learning method with CNN and with Generative Adversarial Network (GAN) are applied to solve SR problem. However, the results are not good and always have distortions. In order to get a better High-Resolution result, an end-to-end method called FSR-Net combines face prior with CNN is proposed and performed well on face SR task. In this project, we are going to re-implement the structure and test the performance.

1. Introduction

Face Super-Resolution(SR) is the task to recover the high-resolution (HR) face image from a low-resolution(LR) input face image. This topic is very important since many other face-related tasks like face-parsing and face recognition can do much better with HR rather than LR images.

Face SR is special from general image SR tasks. It can facilitate face prior knowledge which contains a lot of information about face. Face priors like facial correspondence field is useful for recover face shape [16], and facial components information gives details of face image [11]. There are some methods using face priors and reaches good performance. However, all the methods are multi-stage but not end-to-end [4]. To solve the face SR problem with face prior and an end-to-end structure, Chen et al. proposed FSR-Net [4].

Face Super-Resolution Network (FSR-Net) estimates both facial landmark heatmaps and parsing maps during training process and uses the estimated prior with CNN to generate HR face image. The end-to-end structure ensures that the weight for prior estimation and HR image generation can be learned concurrently. The FSR-Net contains two main parts. The first part is the coarse SR network which generates a low quality HR image. This step is indispensable because estimate prior from a very LR image is hard. The estimated prior will be more precise with a better LR face image. After the coarse SR network, the output will be sent to a fine SR encoder and a prior estimation network simultaneously. Fine SR encoder is responsible for extract image features while the other estimates the face priors. The output of the two are feature maps with same size. Then, we concatenate the two output through the channel dimension and generate the final HR image with the fine SR decoder.

In this work, we mainly re-constructed the whole architecture of FSR-Net and validated the structure with experiments.

2. Related work

In this work, we used face prior in an end-to-end structure. The idea came from other works of multi-stage structure with face prior and end-to-end face SR algorithms.

Facial Prior Facial prior information have been used in many face SR algorithms. Early techniques were based on the assumption that face structure follows a similar setting with small variance. Baker and Kanade [1] tried to learn the spatial distribution of the image gradient for frontal face images. Kolouri et al. [7] fitted the LR image with a non-linear Lagrangian model to generate HR face images. Yang et al. [14] used mapping between facial components to help generate HR image. However, the mapping is produced by landmark detection which is difficult for LR image.

In recent years, people tried to use deep convolutional neural network to solve face SR task. Zhu et al. [16] successfully generate HR image for unaligned LR faces. They do face SR and estimate dense correspondence filed alterna-

[*] equal contribution

tively. Song et al. [11] proposed a method to first generate facial components by CNN and recover the HR face from these components.

End-to-end Training Many end-to-end training method have reached a good performance in face SR task. Ledig et al. [9] proposed Super-Resolution Generative Adversarial Network (SRGAN) that uses perceptual loss function to do photo-realistic image SR. While Yu et al. [15] proposed transformative discriminative auto encoder to deal with unaligned tiny LR images with noise. Furthermore, Cao et al. [2] used deep reinforcement learning to discover attended patches and recover the HR image through exploiting global image interdependency.

3. Data

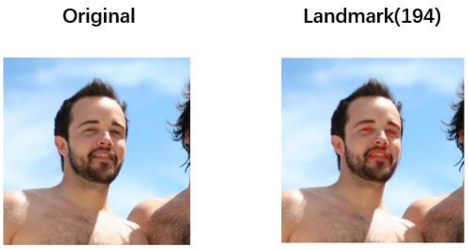

Figure 1. Landmark of human face in Helen dataset.

Our data comes from Helen dataset[8]. It contains 2330 face images with 194 landmark Fig[1] for 11 different face features, such as the eyes, nose, mouth, eyebrows, and jawline. Based on the those 194 landmark, we also have 11 parsing maps Fig[2]. In our experiment, we will use those landmarks and parsing maps as our facial prior.

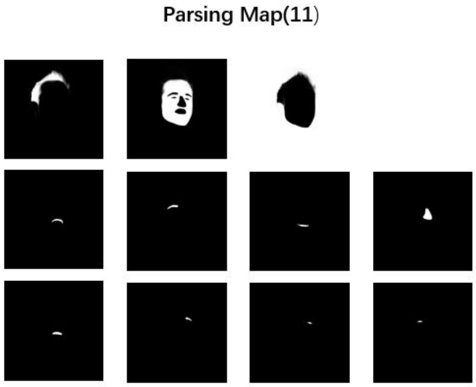

Figure 2. Parsing Map of human face in Helen dataset.

In total 2330 face images, we split into two parts: a training set containing 2280 images and test set containing 50 images. We perform data augmentation on the training images. Specifically, we rotate the original images by 90°, 180°, 270°, and flip them horizontally and vertically. We also crop human face from the original face images. Besides, we add some Gaussian noise to low resolution images after shrinking to 1/8 and enlarging using bicubic interpolation. After applying those data augmentation skills, we could enlarge our training dataset and reduce over-fitting problems in a certain degree.

4. Method

4.1. Overview of FSR-Net

The basic FSR-Net mainly contains four parts: coarse SR network, fine SR encoder, prior estimation network and fine SR decoder. The structure is shown in Figure 3.

To get a better prior estimation, we first use the coarse SR network to generate a low quality HR image. We represent the output of coarse SR network as y_c. Then, y_c is sent to both the fine SR encoder and the prior estimation network. After these two network, we can get the feature extracted by the encoder and the estimated prior. We denote the output as f and p respectively. Then, we concatenate f and p through the channel dimension and input to the final fine SR decoder to generate the HR image y.

Suppose the ground truth for the HR image and the face prior is given by $\hat{y}$ and $\hat{f}$, the loss of FSR-Net is

$$\mathcal{L}_F(\Theta) = \frac{1}{2N}\sum_{i=1}^{N}\{||\hat{y}^{(i)} - y_c^{(i)}||^2 + ||\hat{y}^{(i)} - y^{(i)}||^2 + \lambda||\hat{p}^{(i)} - p^{(i)}||^2\} \tag{1}$$

where Θ denotes the parameter set, N is the number of samples in a batch, i represents the ith sample and λ is the weight for prior loss.

4.2. Detailed Structure

4.2.1 Coarse SR Network

The coarse SR network aims to generate a coarse HR image from the original LR image. It is needed because the difficulty of estimating face prior from the LR image. The coarse SR network is set to reduce the difficulty. The structure of the coarse SR network is shown in Figure 3. It begins with a 3×3 convolution layer and followed by 3 residual blocks [6] and finally followed with another 3×3 convolution layer. The output is the generated coarse HR image.

4.2.2 Fine SR Network

In the fine SR network, the coarse HR image is sent to two branches, the fine SR encoder and the prior estimation net-

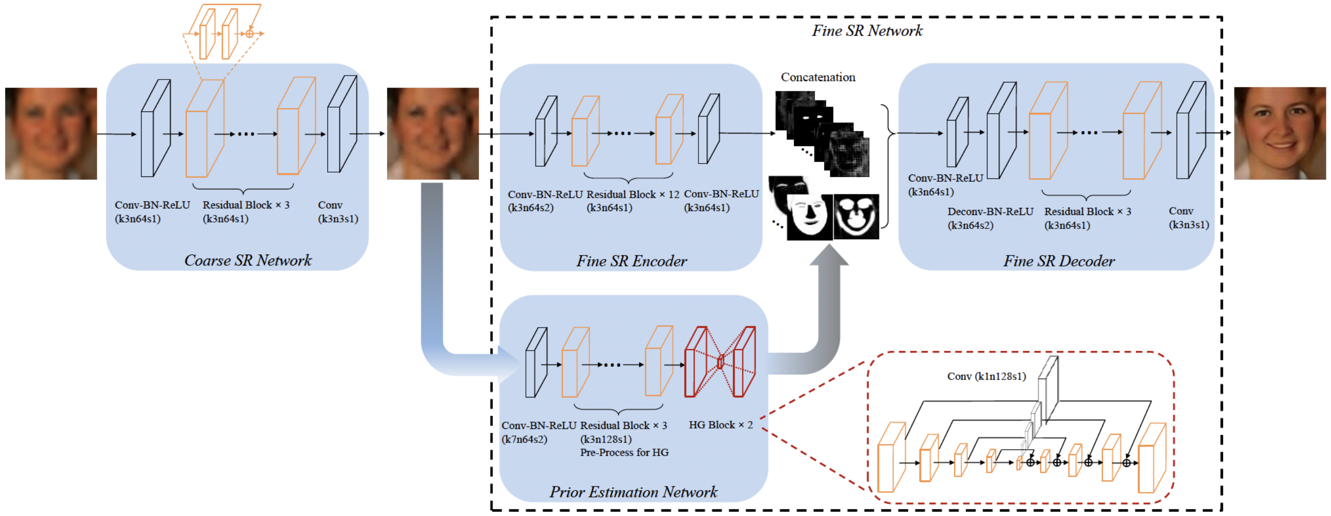

Figure 3. Network structure of FSR-Net. k3n64s1 indicates the kernel size 3*3, the channel length 64 and the stride 1.

work to extract face features and estimate facial priors respectively. Then the decoder uses two results together to generate the final HR image.

Prior Estimation Network There are two types of face priors, the facial shape and texture. In this project, we choose to leverage the shape prior for face. There are two reasons. First, shape information for a face can be well preserved in a LR image compared to texture. Second, representing shape prior is much easier than texture prior. We can use areas or points to describe the shape information. However, how to represent the texture is not clear. In FSR-Net, we used face parsing and landmark heatmaps as the prior. Face paring estimates the segmentations of different face components, and landmarks provide the accurate locations of facial keypoints.
For prior estimation, we adopt the Hour-Glass (HG) structure [10]. The detailed structure of the prior estimation network is shown in Figure 3. To preserve spatial information and features with different scales, skip connection between symmetrical layers is applied in the hourglass block. An 1×1 convolution layer is followed to process the obtained features. Finally, the hour glass output is connected to two separate 1×1 convolution layers to generate the landmark heatmaps and the parsing maps.

Fine SR Encoder In fine SR Encoder, we used the residual blocks to do feature extraction. Considering the computation cost, we tried to down-sample the inputs. The fine SR encoder starts with a 3×3 convolutional layer of stride 2 to down-sample the size to 64×64. Then the features can be extracted by ResNet structure.

Fine SR Decoder The fine SR decoder uses two results together to generate the final HR image. We concatenated the prior feature p and image feature f through the channel dimension as the input of the decoder. A 3×3 convolutional layer is used to reduce the number of feature maps to 64. Then, a 4×4 deconvolutional layer up-sampled the feature map to size 128×128. To decode the features, we used 3 residual blocks. Finally, a 3×3 convolutional layer is followed to generate the HR image.

5. Experiments

5.1. Implementation Details

Dataset We basically conduct our experiment on Helen dataset[8]. Experimental setting for Helen dataset is described in previous section Sec[3]. We split whole helen dataset 2330 images into two parts randomly, 2280 images for training and 50 images for testing. Beyond that, we use other 100 face images to evaluate our result. The original 100 face images are high resolution and we use the same way to preprocess high resolution images and get corresponding low resolution images to run evaluation.

Data Augmetation Data augmentation is a good way to enlarge our limited dataset and reduce over-fitting problems in some degree. The data augmentation method is described in previous section Sec[3]. Our training set is augmented through rotation by 90°, 180°, 270°, flipping horizontally and vertically. Besides, we also add some Gaussian noise to low resolution images after shrinking to 1/8 and enlarging using bicubic interpolation. By applying this data augmentation method, we can reduce large amount of over-fitting brought by the deep neural network.

Training Setting After preparation of our data, we start training our model with our own computer with a NVIDIA

GeForce GTX 1050Ti. The model is trained used the Adam Optimizer Algorithm (which is different from FSRNet [4]) with an initial learning rate 2.5×10^{-4}, and the mini-batch size of 2 due to the limitation of our GPU memory size. On the server provided by this course, we increased the mini-batch to 8. We simply use the same hyper-parameter as FSRNet [4], setting λ = 1, $\gamma_{\mathbf{C}} = 10^{-3}$ and $\gamma_{\mathbf{P}} = 10^{-1}$. Training our model on Helen dataset takes $\sim$8 hours on GeForce GTX 1050Ti GPU.

Code To construct our model, we basically refer to the code structure provided by another paper, SRNet [12], which proposed a deblur model. We use the code structure from what they posted on GitHub, `https://github.com/jiangsutx/SRN-Deblur/`, and substitute data loading method and the core network structure.

5.2. Result

After we get our model after training, we tried to get high resolution image from low resolution image. We ran our model on the evaluation data and get the result shown in Fig [4].

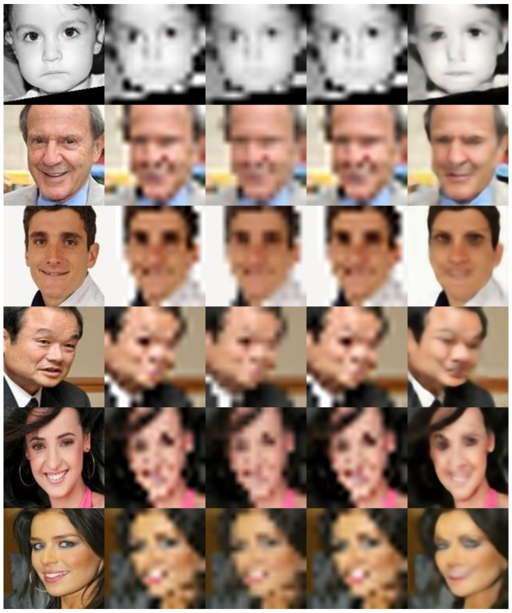

Figure 4. Reconstructed High Resolution images on evaluation dataset. The first column is original high resolution image; the second column is bicubic low resolution image; the third column is bilinear low resolution image; the fourth column is a sample image reconstruct through gaussian filter; the last column is our reconstructed high resolution image.

If we look into the details between the low resolution image and reconstructed high resolution image Fig [5], we can find that we extract the boundary information and human face information. The boundary in low resolution image is much clear than the bicubic low resolution image. And we can also see that there are more details in human's eyes and eyebrows. The eye and eyebrows can be distinguished in reconstructed high resolution image instead of an obscure colored block.

Method	MSE	SSIM	PSNR
cubic	594.3	0.5901	20.67
bilinear	557.5	0.5986	20.95
gaussian	588.3	0.5910	20.72
Our model	308.69	0.6645	23.51

Table 1. Error and Score of three metrics.

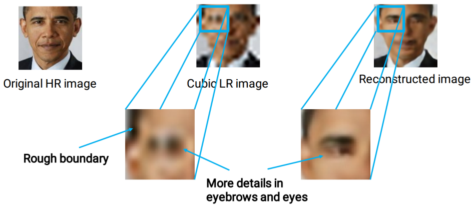

Figure 5. Detailed Comparison between low solution image and reconstructed high resolution image.

5.3. Comparison

Since we didn't have enough time to apply other state-of-art super-resolution algorithms on this face super resolution task, we only compared our results with low resolution image and other image get through sample filter (Gaussian Filter). We used three metrics to evaluate our model performance: Peak Signal to Noise Ratio (PSNR), Structural Similarity Index (SSIM) and Mean Squared Error (MSE). Combining those three metric, we can reflect our model performance more comprehensively. Notice that the error MSE bigger, the performance is worse while SSIM and PSNR score higher, the performance is better. The score of those three metric is shown in Table [1].

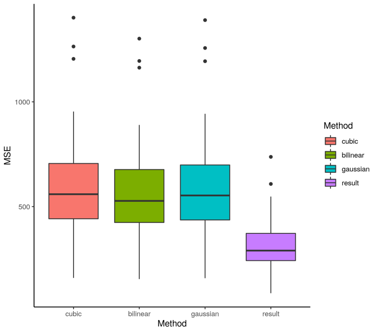

Figure 6. Boxplot of MSE error on evaluation data.

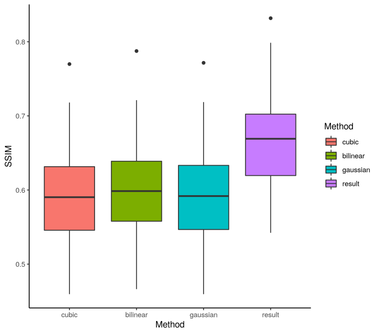

Figure 7. Boxplot of SSIM score on evaluation data.

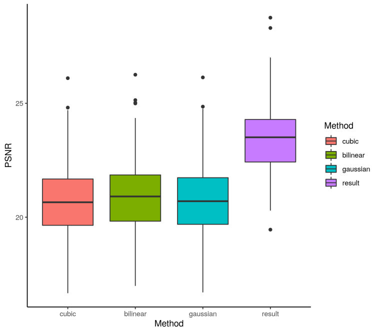

Figure 8. Boxplot of PSNR score on evaluation data.

As you can see in Fig[6], Fig[7] and Fig[8], the model get lowest MSE error and highest score in both SSIM and PSNR. Although we only compared with simple methods, it could prove that our model functions on face super-resolution tasks.

6. Conclusion

Basically, we successfully re-implemented the face super resolution algorithm provided FSRNet [4] on helen dataset. We also did some other data augmentation methods and switch to other optimizer. We also tried some other method to improve the original FSRNet, but none of them were satisfying. However, our trained model indeed functions on face super-resolution task and could extract information from the low resolution image. Our experiment could prove that, facial prior can better guide the facial super-resolution task. Facial prior is important to reconstruct face image. Some other face super resolution algorithms [5] that only find the relationship between low resolution image and high resolution image may distort human face and could not guarantee the performance. With facial prior, the network would learn human face information and reconstruct the face more precisely. In this way, no more distort would happen on reconstructed high resolution image. Beyond that, it is also a good attempt to combine image information and other information together. In the future, computer vision may only rely on a combination of image and other information instead of only on a branch of pixels.

7. Future work

Due to time limitation, we still have a lot of work to do. The first thing we could improve is on network structure. From our result, we can see that we still have a difference with the original high resolution image, especially in details. Thus, we can apply local branch and global branch strategy to not only focus on the image from a high level, but also reconstruct more details on human faces. In our experiment, we didn't implement the GAN part in FSR-Net [4]. The discriminator in GAN part can improve the realness of reconstructed picture. Also, we use the new structure deepLabv3 [3] to subtitle the basic U-Net structure. Secondly, other usage of facial prior could be more helpful. Current network only extract facial prior information to some channel and then reconstruct. From attention mechanism [13], using facial prior as attention layer maybe another way. Also, the current model couldn't have same performance on other scales. If the scale between the high resolution and low resolution is much higher than ×8, the picture would still be obscure. We need also consider applicability for different scales in the future.

References

[1] S. Baker and T. Kanade. Hallucinating faces. *fg*, 2000:83–88, 2000.

[2] 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*, pages 690–698, 2017.

[3] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *ECCV*, 2018.

[4] Y. Chen, Y. Tai, X. Liu, C. Shen, and J. Yang. Fsrnet: End-to-end learning face super-resolution with facial priors. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[5] Z. Chen and Y. Tong. Face super-resolution through wasserstein gans. *CoRR*, abs/1705.02438, 2017.

[6] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[7] S. Kolouri and G. K. Rohde. Transport-based single frame super resolution of very low resolution face images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4876–4884, 2015.

[8] V. Le, J. Brandt, Z. Lin, L. Bourdev, and T. S. Huang. Interactive facial feature localization. In A. Fitzgibbon, S. Lazebnik, P. Perona, Y. Sato, and C. Schmid, editors, *Computer Vision – ECCV 2012*, pages 679–692, Berlin, Heidelberg, 2012. Springer Berlin Heidelberg.

[9] C. Ledig, L. Theis, F. Huszár, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, Z. Wang, et al. Photo-realistic single image super-resolution using a generative adversarial network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4681–4690, 2017.

[10] A. Newell, K. Yang, and J. Deng. Stacked hourglass networks for human pose estimation. In *European conference on computer vision*, pages 483–499. Springer, 2016.

[11] Y. Song, J. Zhang, S. He, L. Bao, and Q. Yang. Learning to hallucinate face images via component generation and enhancement. *arXiv preprint arXiv:1708.00223*, 2017.

[12] X. Tao, H. Gao, X. Shen, J. Wang, and J. Jia. Scale-recurrent network for deep image deblurring. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[13] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, and I. Polosukhin. Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 5998–6008. Curran Associates, Inc., 2017.

[14] C.-Y. Yang, S. Liu, and M.-H. Yang. Structured face hallucination. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1099–1106, 2013.

[15] X. Yu, B. Fernando, B. Ghanem, F. Porikli, and R. Hartley. Face super-resolution guided by facial component heatmaps. In *The European Conference on Computer Vision (ECCV)*, September 2018.

[16] S. Zhu, S. Liu, C. C. Loy, and X. Tang. Deep cascaded bi-network for face hallucination. In *European conference on computer vision*, pages 614–630. Springer, 2016.