SF-SRGAN: PROGRESSIVE GAN-BASED FACE HALLUCINATION

M. N. Favorskaya [1,*], A. I. Pakhirka [1]

[1] Reshetnev Siberian State University of Science and Technology, 31 Krasnoyarsky Rabochy ave., Krasnoyarsk, 660037 Russian Federation - (favorskaya, pahirka)@sibsau.ru

Commission II, WG II/8

KEY WORDS: Face Hallucination, Super-resolution, Generative Adversarial Network, Progressive Upsampling, Identity Loss, Image Enhancement.

ABSTRACT:

Facial hallucination is a technique that has emerged recently thanks to advances in deep learning. It can be used in various tasks such as face recognition in the wild, human identification, pedestrian re-identification, face analysis, and so on. We propose a wavelet-integrated trained face hallucination model to synthesize photorealistic face images called SF-SRGAN. The multi-stage progressive hallucination strategy is based on GAN architecture. The proposed generator consists of sequential cascade modules, each of which increases the scale by 2×. Each module has a complex structure of two branches: a progressive face hallucination branch for feature extraction and reconstruction and edge-preserving branch for high frequency detail extraction. The main difference from other progressive GAN-based face hallucination networks is that the two branches fuse followed by each cascade 2×. The model is trained and tested on popular public face datasets such as the CelebA-HQ dataset, the LFW dataset, and the Helen dataset with promising photorealistic results.

1. INTRODUCTION

Face hallucination belongs to a family of single image super-resolution (SISR) techniques aimed at generating super-resolution (SR) images from observed low-resolution (LR) facial input images. This method is used to improve face images in various tasks such as face recognition, human identification, pedestrian re-identification, face analysis, and so on. Moreover, the facial images captured by real surveillance cameras have very low resolution, and the SISR application enhances perception and visibility. SISR is an ill-posed problem since multiple SR images can be generated from the same LR image. Generally speaking, face hallucination methods can be divided into interpolation-based (for example, bicubic interpolation), statistics-based and learning-based approaches. In recent decades, the learning-based approaches have dominated due to higher upscaling factors compared to the first two types (Jiang et al., 2023).

Learning-based methods of face hallucination are divided into three main categories:

- Global face hallucination methods are robust to noise, but due to inaccurate reconstruction, the edges of the face are often blurred (Shi et al., 2020).
- Local face hallucination methods demonstrate better representation ability based on dividing the image into overlapping small patches (Lu et al., 2018). For this, various regression models are used.
- Two-stage facial hallucination methods attempt to combine global and local information by taking advantage of the two approaches mentioned above (Jiang et al., 2022).

Another criterion for taxonomy is the application of traditional machine learning and deep learning models. Deep learning models use an end-to-end mapping function to learn SISR paradigms in a view of convolutional neural network (CNN) and generative adversarial network (GAN) representations, although recurrent neural networks are sometimes applied. It should be noted that GAN-based methods generate more realistic face images compared to CNN-based methods, which use pixel-wise loss and generate smooth face images. They also provide an increased scaling factor in the SISR problem. GAN-based methods are classified into general GAN-based methods, which are learned from paired or unpaired data, and generative prior-based methods, in other words, pre-trained generative models (Jiang et al., 2023).

Progressive face super-resolution is a sub-branch of the SR domain for face image reconstruction. On the one hand, the main challenge of the face SISR problem is to restore essential features of the face without distortion. On the other hand, due to their nature, the GAN-based models that generate photo-realistic 4× or 8× super-resolved face images can add fake features.

In this study, we propose a wavelet-integrated trained face hallucination approach to synthesize photorealistic face images under assumption that the LR images are significantly degraded by noise, blur and JPEG compression. The main idea is to use a digital wavelet transform (DWT) branch in each stage of 2× image upscaling. We tested our approach on popular public face datasets such as the CelebA-HQ dataset, the LFW dataset, and the Helen dataset with promising results.

The rest of the paper is structured as follows. Section 2 discusses related work. Section 3 provides the background to

* Corresponding author

the face hallucination. The proposed approach is presented in Section 4. The details of the experiment are covered in Section 5. Section 6 concludes the paper.

2. RELATED WORK

The appearance of the first deep architectures for the SISR problem, such as collaborative auto-encoders, deep CNNs and deconvolutional networks, dates back to the mid-2010s (Dong et al., 2016). One of the pioneering works was the study of implementing a sparse coding based model in a cascade end-to-end architecture (Wang et al., 2015). The sparse coding based network used the bicubic-upscaled LR image in the cascade that had the same scaling factor *s*. Three cascades have been proposed with parameters scale×1, scale×*s* and scale×s^2. This approach was later called progressive multi-scale design principle. To obtain photorealistic results, the ProGanSR model was proposed in (Wang et al., 2018). These authors demonstrated that applying a GAN model can improve the reconstruction quality for all upsampling factors simultaneously and can scale to high upsampling factors (scale×8). At the intermediate stages, a 2× upsampling of the input from the previous level was sequentially performed. To simplify the data propagation within the network, an asymmetric pyramidal structure was developed with a large number of layers at the lower levels. Some studies use the principle of Laplacian pyramids, when each level learns to predict a residual between a simple upscale of the previous level and the desired result (Lai et al., 2019; Anwar et al., 2022). Another branch of study is how to prevent the tendency of GAN-based approaches to add fake textures and even artifacts to make the SR images look naturally attractive. To this end, a super-resolution method was proposed through a perceptually oriented progressive network with three modules reconstructing content, structure and photo-realism (Hui et al., 2021).

Progressive face super-resolution is a sub-branch of the SR domain for face image reconstruction. On the one hand, the main challenge of the face SISR problem is to restore essential features of the face without distortion. On the other hand, due to their nature, the GAN-based models that generate photo-realistic 8× super-resolved face images can add fake features In (Kim et al., 2019), it was shown how facial landmarks help to generate SR facial images with fully preserved facial details using progressive training, facial attention loss, and distillation of face alignment network. In (Zhang et al., 2022), the task of hallucinating an authentic high-resolution face from an occluded thumbnail has been studied. The occluded and tiny face images at a resolution 16×16 pixels were the inputs to a multi-stage progressive upsampling and inpainting GAN (Pro-UIGAN) that upscaled the resolution to 8×. Pro-UIGAN was based on a multi-stage progressive hallucination strategy and included Pro-UInet for creating facial landmark heatmaps, as well as the Local-D and Global-D modules that made the hallucinated facial images photorealistic. A progressive upsampling based cascaded recurrent convolutional network was proposed in (Liu et al., 2021). However, recursive learning has its advantages and disadvantages, such as learning long-term context dependencies and vanishing or exploding gradient, respectively.

3. BACKGROUND

Face hallucination is the inverse problem recovering the SR image from the LR image I_{LR} using the following equation:

$$I_{SR} = F\left(I_{LR}, \delta\right), \tag{1}$$

where I_{SR} = the super-resolution image
I_{LR} = the low-resolution image
F = the inverse degradation model
δ = the parameters of F

The optimization of δ is defined by the following equation:

$$\hat{\delta} = \arg\min_{\delta} L\left(I_{SR}, I_{HR}\right), \tag{2}$$

where L = the loss between I_{SR} and I_{HR}
I_{HR} = the high-resolution image
$\hat{\delta}$ = the optimal parameter of the trained model

However, the degradation model F and its parameters δ are unknown in the real environment. We are only dealing with LR images. A common way to evaluate a model is to simulate HR images degradation process by generating LR images and using pairs of LR and HR images to train the model. For this, operators such as subsampling, blurring, noise, or even JPEG compression are used.

It should be noted that the difference between an SR image and an HR image is in that an SR image is reconstructed with any method, while an HR image is an image captured by a camera and is usually included in the dataset (with or without a corresponding LR image).

4. PROPOSED APPROACH

The multi-stage progressive hallucination strategy involves step by step tuning from coarse results to fine results. These steps depend on the end goal of recognizing or improving the LR image and can be burdened by occlusions and facial position. In addition, the complex mapping of an LR image to an SR image is easier to implement in steps.

The classic SR approach is to use an enhanced super-resolution generative adversarial network (ESRGAN) (Wang et al., 2018) or its later version, the Real-ESRGAN model (Wang et al., 2021a). The Real-ESRGAN model can simulate complex real world degradations by modifying the ESRGAN generator (employing the pixel-unshuffle layer) and the ESRGAN discriminator (using the U-Net design with skip connections). Recently, new original approaches have been proposed. The SwinIR model for image restoration based on the Swin Transformer consists of three parts: shallow feature extraction, deep feature extraction and high-quality image reconstruction (Liang et al., 2021). High-resolution image synthesis based in latent diffusion models was developed for image inpainting and class-conditional image synthesis (Rombach et al., 2022). In addition, several restoration methods can be applied to enlarged images to improve the visibility of the face (so called hallucination of the face). Thus, generative facial prior (GFP) was incorporated in the GAN-based face restoration process to achieve a good balance of realness and fidelity (Wang et al., 2021b). Codebook lookup transformer called CodeFormer modeled the global composition and context of the low-quality faces (Zhou et al., 2022).

The proposed solution with the wavelet module was motivated by the preliminary experimental results of face reconstruction presented in Table 1. As shown in Table 1, mean PSNR values

and mean SSIM values calculated for nearly 1,000 face images from the LFW, CelebA-HQ and Helen datasets are close when reconstructing LR images with low degradation. As can see from Table 1, additional reconstruction models such as CodeFormer and GFP-GAN reduce the mean PSNR values and mean SSIM values, making face images more realistic. However, these values fail when the LR images are significantly degraded by noise, blur, or JPES compression.

Model	Mean PSNR, dB	Mean SSIM
ESRGAN	24.54	0.82
ESRGAN + CodeFormer	24.10	0.81
ESRGAN + GFP-GAN	25.67	0.85
R-ESRGAN	27.97	0.88
R-ESRGAN + CodeFormer	25.44	0.83
R-ESRGAN + GFP-GAN	27.39	0.87
SwinIR	27.76	0.88
SwinIR + CodeFormer	25.40	0.83
SwinIR + GFP-GAN	27.28	0.87
LDSR	28.45	0.88
LDSR + CodeFormer	25.67	0.83
LDSR + GFP-GAN	27.85	0.87

Table 1. Mean values of face image reconstruction.

Possible schemes for face hallucination from simple and complex to the proposed one are depicted in Figure 1. At each stage (Figure 1c), the input from the Pre-processing module is fed into the Upscaling module and the DWT module. The Upscaling module has a GAN architecture, in our case ESRGAN. In order to improve structural information, we use an approach similar to the DWSR network (Guo et al., 2017), when, first, the input spatial image is transformed by a direct DWT, second, DWSR network upsamples the DWT coefficients of the LR image to the DWT coefficients of SR scale, and, third, the upsampling map is transformed by the inverse DWT and tuned for the upsampled spatial image. The proposed architecture is called spatio-frequency SRGAN (SF-SRGAN).

The proposed generator consists of sequential cascade modules, each of which increases the scale by 2×. In our experiments, scaling factors 4× and 8× were used. Each module has a complex structure of two branches: a progressive face hallucination branch for feature extraction and reconstruction and edge-preserving branch for high frequency detail extraction. The main difference from other progressive GAN-based face hallucination networks is that the two branches fuse followed by each cascade 2×. This fused module provides edge-preserving upsampling, not just the usual landmarks for facial reconstruction.

The progressive face hallucination branch includes a hierarchical feature extraction unit and an upsampling unit. The hierarchical feature extraction is motivated by the fact that elements like nose, eyes, lips, ears and wrinkles are local attributes in an image and should be extracted by convolution layers with small kernel sizes such as 1×1 and 3×3. At that time, the global texture features are extracted using larger kernel sizes 5×5. The LeakyReLU activation function introduces non-linearity to the branch. The final feature map is then obtained by concatenating the feature maps from each convolution layer with a different kernel size. The reconstruction is implemented using a deconvolution layer to upsample feature maps.

5. EXPERIMENTS

The proposed SF-SRGAN model was trained and tested on popular public face datasets: LFW, CelebA-HQ, and Helen. The LFW dataset (Huang et al., 2007) provides 13,233 images at 250×250 resolution of 5,749 individuals under different conditions collected from the web. For LFW dataset, 11,380 images and the remaining 1,853 images are used for training and evaluation, respectively. The CelebA-HQ dataset (Lee et al., 2020) consists of 30,000 face images at 1024×1024 resolution under various poses, expressions, and backgrounds. We use 28,432 images are for training, and 1,568 images are for testing. The Helen dataset (Le et al., 2012) is composed of 2,330 in-the-wild face images with labelled facial components, e.g., eyebrows, lips, nose, skin, hairs, etc. For Helen dataset, there are 2,280 images for training and 50 images for testing. For image augmentation, only a random horizontal flip was performed in all experiments.

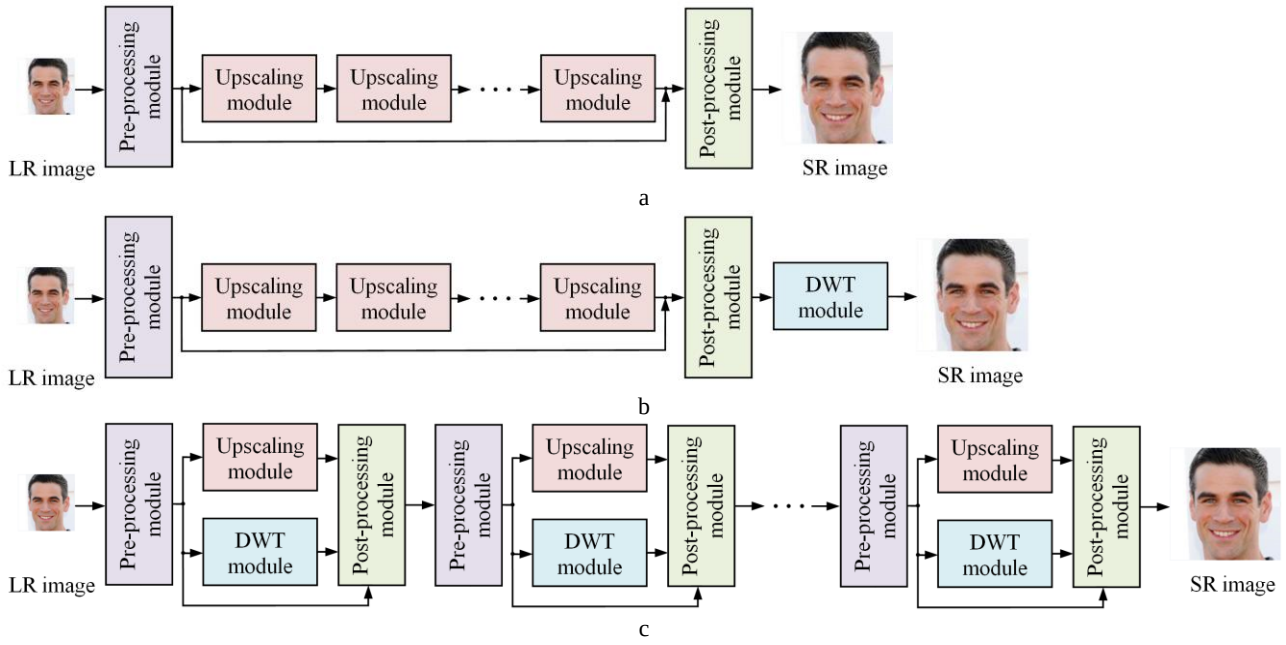
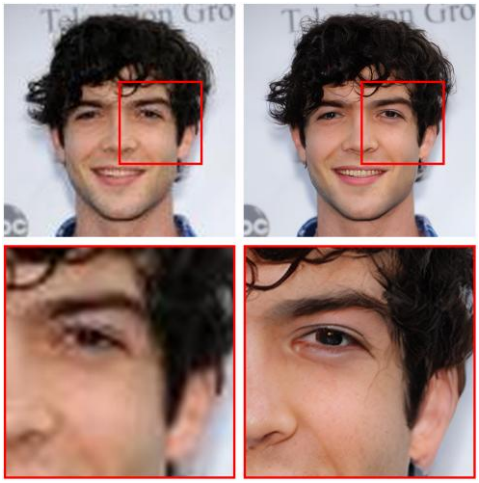


Figure 1. Schemes for face hallucination: a) classical progressive approach, b) classical progressive approach with refined DWT, c) proposed SF-SRGAN model.

We used peak signal to noise ratio (PSNR) and structural similarity index (SSIM) as standard evaluation metrics. However, high PSNR valur of the resulting image does not guarantee a high visual quality (Figure 2). Bicubic interpolation may have a higher PSNR value, but the result of GAN-based models is characterized by better human visibility. Therefore, learned perceptual image patch similarity (LPIPS) evaluation metric are additionally used. The LPIPS metric is often used to measure the perceptual similarity between two images (Zhang et al., 2018). At the same time, the SSIM metric makes it possible to evaluate structural similarity.

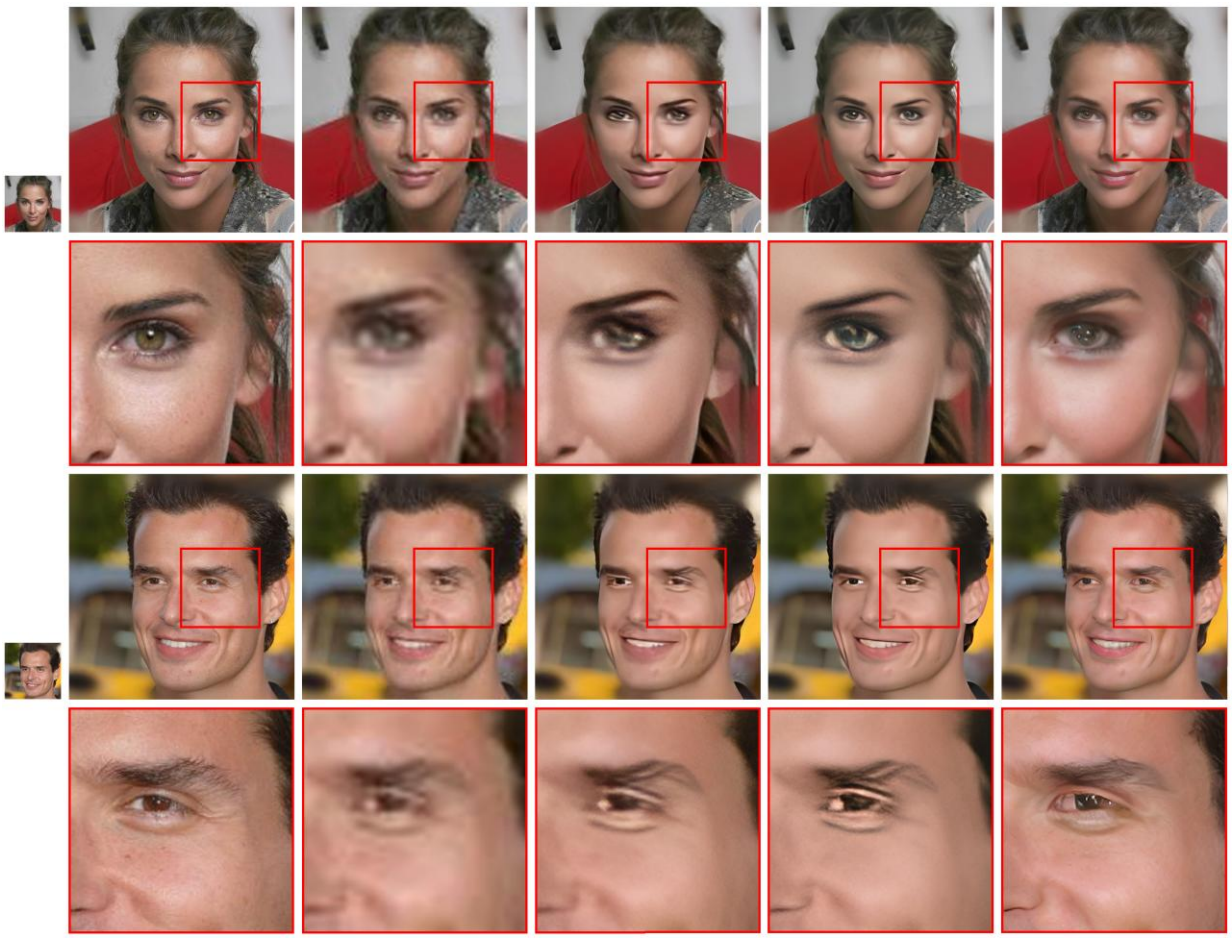
Preliminary experiments have revealed difficulties in obtaining high-quality SR facial images (see Figure 3), which requires further experimental studies with advanced GAN-based architecture and improvement of the learning process.

The proposed network model was implemented on Python using the Pytorch repository. The GPUs used in experiment NVIDIA Geforce RTX 2080 Ti (11GB). The operating system is MS Windows 10. In generator, the number of RRDB (residual in residual dense block) is chosen as 12. Proposed GAN-based model is optimized with Adam. During the training, the batch size is set to 8. Table 2 shows the results of face reconstruction. Higher scores of the LPIPS metric mean more differences, while the lower scores of the LPIPS metric mean more similarities.



a b

Figure 2. Example of 4× image upscaling with the increased fragments with: a) bicubic interpolation (PSNR = 31.38, LPIPS = 0.174), b) proposed SF-SRGAN model (PSNR = 29.06, LPIPS = 0.112).



a b c d e f

Figure 3. Examples of face hallucination with the increased fragments: a) input LR images with 128×128 resolution, b) original HR images with 512×512 resolution, c) 4× image upscaling with bicubic interpolation, d) 4× image upscaling with the ESRGAN model, e) 4× image upscaling with the ESRGAN + DWSR model, f) 4× image upscaling with the proposed SF-SRGAN model.

Model	Mean PSNR, dB	Mean SSIM	Mean LPIPS
ESRGAN	24.54	0.82	0.175
R-ESRGAN	27.97	0.88	0.160
SwinIR	27.76	0.88	0.119
LDSR	28.45	0.88	0.108
ESRGAN+DWSR	26.87	0.89	0.102
SF-SRGAN	27.95	0.89	0.096

Table 2. Mean values of face image reconstruction with the LPIPS metric.

The proposed SF-SRGAN model can generate more realistic face images. This approach was developed to increase the sensitivity to the details, which is not enough in CNN models and eventually leads to over-smooth result. The sub-band images in the wavelet domain represent low-frequency global topology and high-frequency textures, respectively, which improve the quality of super-resolution images.

6. CONCLUSIONS

This paper proposes a wavelet-integrated trained face hallucination model to synthesize photorealistic facial images called SF-SRGAN. Texture loss is a common problem of existing super-resolution methods. Thus, the facial features are concatenated with the wavelet features to provide a more detailed feature map. This technique is used to compensate the smoothed nature of super-resolution methods. Future modifications can be made to improve the performance of the SF-SRGAN model.

REFERENCES

Anwar, S, Barnes, N., 2022: Densely residual Laplacian super-resolution. *IEEE Trans. Pattern Anal. Mach. Intell.* 44(3), 1192-1204. doi.org/10.1109/TPAMI.2020.3021088.

Dong, C., Loy, C.C., He, K., Tang, X., 2016: Image super-resolution using deep convolutional networks. *IEEE Trans. Pattern Anal. Mach. Intell.* 38(2), 295-307. doi.org/10.1109/TPAMI.2015.2439281.

Guo, T., Mousavi, H.S., Vu, T.H., Monga, V. (2017) Deep wavelet prediction for image super-resolution. *2017 IEEE Conf. Computer Vision and Pattern Recognit. Workshops (CVPRW)*, IEEE, Honolulu, HI, USA, pp. 104-113. doi.org/10.1109/CVPRW.2017.148.

Huang, G.B., Ramesh, M., Berg, T., Learned-Miller, E. 2007: Labeled Faces in the Wild: A Database for Studying Face Recognition in Unconstrained Environments. University of Massachusetts, Amherst, Technical Report 07-49.

Hui, Z., Li, J., Gao, X., Wang, X., 2021: Progressive perception-oriented network for single image super-resolution. *Inf. Sci.* 546, 769-786. doi.org/10.1016/j.ins.2020.08.114.
Jiang, J., Wang, C., Liu, X., Ma, J., 2023: Deep learning-based face super-resolution: A survey. *ACM Computing Surveys* 55(1), Article No. 13, 1-36. doi.org/10.1145/3485132.

Jiang, K., Wang, Z., Yi, P., Lu, T., Jiang, J., Xiong, Z., 2022: Dual-path deep fusion network for face image hallucination. *IEEE Trans. Neural Netw. Learn. Syst.* 33(1), 378-391. doi.org/10.1109/TNNLS.2020.3027849.

Kim, D., Kim, M., Kwon, G., Kim, D.-S., 2019: Progressive face super-resolution via attention to facial landmark. *British Machine Vision Conference (BMVC'19)*, paper no. 192, pp. 1-12. Cardiff, UK.

Lai, W.-S., Huang, J.-B., Ahuja, N., Yang, M.-H., 2019: Fast and accurate image super-resolution with deep Laplacian pyramid networks. *IEEE Trans. Pattern Anal. Mach. Intell.* 41(11), 2599-2613. doi.org/10.1109/TPAMI.2018.2865304.

Le, V., Brandt, J., Zhe, L., Bourdev, D., Huang, T., 2012: Interactive facial feature localization. In: Fitzgibbon, A., Lazebnik, S., Perona, P., Sato, Y., Schmid, C. (eds) *Computer Vision – ECCV 2012 (ECCV 2012)*. LNCS, vol. 7574. *Springer*, pp. 679-692. Berlin, Heidelberg. doi.org/10.1007/978-3-642-33712-3_49.

Lee, C., Liu, Z., Wu, L., Luo, P., 2020: MaskGAN: Towards diverse and interactive facial image manipulation. *The IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR'20)*. IEEE, 5549-5558. Seattle, WA, USA. doi.org/10.1109/CVPR42600.2020.00559.

Liang, J., Cao, J., Sun, G., Zhang, K., Van Gool, L., Timofte, R., 2021: SwinIR: Image restoration using Swin transformer. *2021 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, IEEE, Montreal, BC, Canada, Oct. 11 2021 to Oct. 17 2021, pp. 1833-1844. doi.org/10.1109/ICCVW54120.2021.00210.

Liu, S., Xiong, C., Shi, X., Gao, Z., 2021: Progressive face super-resolution with cascaded recurrent convolutional network. *Neurocomputing* 449, 357-367. doi.org/10.1016/j.neucom.2021.03.124.

Lu, T., Chen, X., Zhang, Y., Chen, C., Xiong, Z., 2018: SLR: Semi-coupled locality constrained representation for very low resolution face recognition and super resolution. *IEEE Access* 6, 56269-56281. doi.org/10.1109/ACCESS.2018.2872761.

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B., 2022: High-resolution image synthesis with latent diffusion models. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, IEEE, New Orleans, LA, USA, 18–24 June 2022, pp. 10674-10685. doi.org/10.1109/CVPR52688.2022.01042.

Shi, Y., Li, G., Cao, Q., Wang, K., Liang Lin, L., 2020: Face hallucination by attentive sequence optimization with reinforcement learning. *IEEE Trans. Pattern. Anal. Mach. Intell.* 42(11), 2809-2824. doi.org/10.1109/TPAMI.2019.2915301.

Wang, Z., Liu, D., Yang, J., Han, W., Huang, T., 2015: Deep networks for image super-resolution with sparse prior. *2015 IEEE International Conference on Computer Vision (ICCV)*, Santiago, Chile, pp. 370-378. doi.org/10.1109/ICCV.2015.50.

Wang, X., Yu, K., Wu, S., Gu, J., Liu, Y., Dong, C., Qiao, Y., Loy, C.C., 2019: ESRGAN: Enhanced super-resolution generative adversarial networks. In: Leal-Taixé, L., Roth, S. (eds) *Computer Vision – ECCV 2018 Workshops. ECCV 2018*. LNCS, vol. 11133. pp. 63-79. Springer, Cham. doi.org/10.1007/978-3-030-11021-5_5.

Wang, Y., Perazzi, F., McWilliams, B., Sorkine-Hornung, A., Sorkine-Hornung, O., Schroers, C., 2018: A fully progressive approach to single-image super-resolution. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, Salt Lake City, UT, USA, pp. 864-873. doi.org/10.1109/CVPRW.2018.00131.

Wang, X., Xie, L., Dong, C., Shan, Y., 2021a: Real-ESRGAN: Training real-world blind super-resolution with pure synthetic data. *2021 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, IEEE, Montreal, BC, Canada, 11-17 October 2021, pp. 1905-1914. doi.org/10.1109/ICCVW54120.2021.00217.

Wang, X., Li, Y., Zhang, H., Shan, Y., 2021b: Towards real-world blind face restoration with generative facial prior. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 20-25 June 2021, Nashville, TN, USA, pp. 9164-9174. doi.org/10.1109/CVPR46437.2021.00905.

Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O. 2018. The unreasonable effectiveness of deep features as a perceptual metric. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 18-23 June 2018, Salt Lake City, UT, USA, pp. 586-595. doi.org/10.1109/CVPR.2018.00068.

Zhang, Y., Yu, X., Lu, X., Liu, P., 2022: Pro-UIGAN: Progressive face hallucination from occluded thumbnails. *IEEE Transactions on Image Processing*, 31, 3236-3250. doi.org/10.1109/tip.2022.3167280.

Zhou, S., Chan, K.C.K., Li, C., Loy, C.C., 2022: Towards robust blind face restoration with codebook lookup transformer. *Thirty-Sixth Conference on Neural Information Processing Systems (NeurIPS 2022)*. New Orleans, Louisiana, USA, 28 November - 9 December 2022, pp. 1-18.