








# Style-Based Attentive Network for Real-World Face Hallucination

Mandi Luo<sup>1,3</sup> , Xin Ma<sup>2</sup> , Huaibo Huang<sup>3</sup> , and Ran He<sup>3</sup>  

<sup>1</sup> School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing, China

<sup>2</sup> Meituan, Beijing, China  
xin.ma@cripac.ia.ac.cn

<sup>3</sup> Institute of Automation, Chinese Academy of Sciences, Beijing, China  
luomandi2019@ia.ac.cn, huaibo.huang@cripac.ia.ac.cn, rhe@nlpr.ia.ac.cn

**Abstract.** Real-world face hallucination is a challenging image translation problem. There exist various unknown transformations in real-world LR images that are hard to be modeled using traditional image degradation procedures. To address this issue, this paper proposes a novel pipeline, which consists of a style Variational Autoencoder (styleVAE) and an SR network incorporated with an attention mechanism. To get real-world-like low-quality images paired with the HR images, we design the styleVAE to transfer the complex nuisance factors in real-world LR images to the generated LR images. We also use mutual information estimation (MI) to get better style information. In addition, both global and local attention residual blocks are proposed to learn long-range dependencies and local texture details, respectively. It is worth noticing that styleVAE is presented in a plug-and-play manner and thus can help to improve the generalization and robustness of our SR method as well as other SR methods. Extensive experiments demonstrate that our method is effective and generalizable both quantitatively and qualitatively.

## 1 Introduction

Single image super-resolution (SISR) aims to infer a natural high-resolution (HR) image from the low-resolution (LR) input. Recently, many deep learning based super-resolution (SR) methods have been greatly developed and achieved promising results. These methods are mostly trained on paired LR and HR images, while the LR images are usually obtained by performing a predefined degradation mode on the HR images, e.g., bicubic interpolation.

However, there is a huge difference between the LR images after bicubic interpolation and real-world LR images. There are various nuisance factors leading to image quality degeneration, e.g., motion blur, lens aberration and sensor noise. Moreover, these nuisance factors are usually unknown and mixed up with each other, making the real-world SR task challenging. The LR generated manually can only simulate limited patterns and methods trained on them inherently lack the ability to deal with real-world SR issues.

---

M. Luo and X. Ma—Contributed equally to this work.

In order to solve this problem, we propose a generative network based on variational autoencoders (VAEs) to synthesize real-world-like LR images. The essential idea is derived from the separable property of image style and image content, which has been widely explored in image style transfer [7, 14, 18]. It means that one can change the style of an image while preserving its content. Based on these previous researches, we propose to consider the fore-mentioned nuisance factors as a special case of image styles. We then design styleVAE to transfer the complex nuisance factors in real-world LR images to generated LR images. In this manner, real-world-like LR images, as well as LR-HR pairs, are generated automatically. Furthermore, styleVAE is presented as a plug-and-play component and can also be applied to existing SR methods to improve their generalization and robustness.

In addition, we build an SR network for real-world super-resolution. Following the principle of global priority in human visual perception systems, our proposed SR network consists mainly of two modules. On the one hand, we develop a global attention residual block (GARB) to capture long-range dependency correlations, helping the SR network to focus on global topology information. On the other hand, we introduce a local attention residual block (LARB) for better feature learning, which is essential to infer high-frequency information in images.

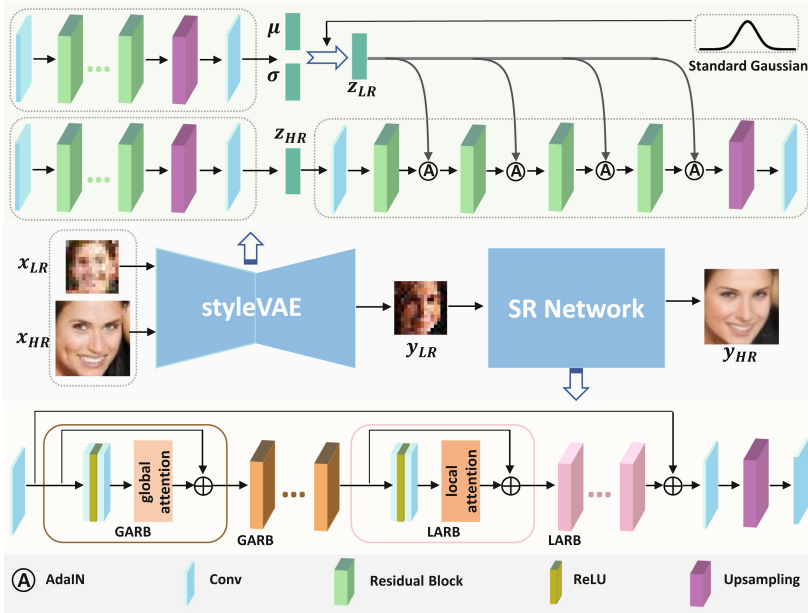
In summary, we make the following contributions: (1) we propose to generate paired LR and HR images with a newly designed styleVAE by learning real-world degradation modes; (2) we propose an SR network with two kinds of attention modules for real-world super-resolution; (3) extensive experiments on real-world LR images demonstrate that styleVAE effectively facilitates SR methods and the proposed SR network achieves state-of-the-art results.

## 2 Related Work

Some previous methods make use of the specific static information of face images obtained by the face analysis technique. Zhu *et al.* [28] utilized the dense correspondence field estimation to help recovering textual details. Meanwhile, some other methods use face image prior knowledge obtained by CNN or GAN-based network. For example, Chen and Bulat [2, 5] utilized facial geometric priors, such as parsing maps or face landmark heatmaps, to super-resolve LR face images. Moreover, some wavelet-based methods have also been proposed. Huang *et al.* [13] introduced a method combined with wavelet transformation to predict the corresponding wavelet coefficients.

## 3 Methodology

Figure 1 shows the overall architecture of our method that consists of two stages. In the first stage, styleVAE is proposed to generate real-world-like LR images. In the second stage, the generated LR images paired with the corresponding HR images are fed into the SR network.



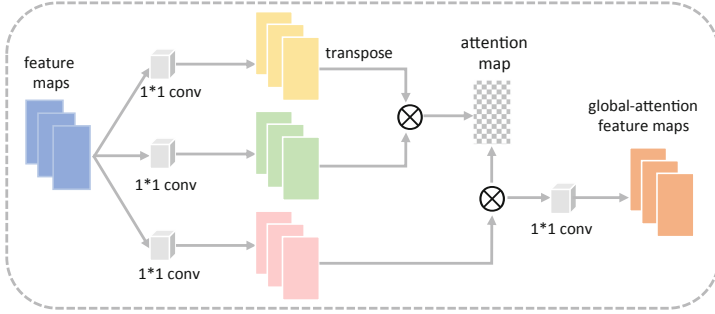
**Fig. 1.** Overall architecture the proposed network. StyleVAE takes unpaired LR and HR images as its inputs to generate real-world-like LR images. SR network takes paired real-world-like LR and HR images as its inputs. By simulating image degradation processes in reality through styleVAE, we can improve the performance of SR methods for real-world SISR.

### 3.1 Style Variational Autoencoder

We adopt Adaptive Instance Normal (AdaIN) to transfer nuisance factors in real-world LR images to generated LR images. There are two inference networks  $E_{LR}$  and  $E_{HR}$ , and one generator  $G$  in styleVAE.  $E_{LR}$  and  $E_{HR}$  project input real-world LR images and HR images into two latent spaces, representing style information and content information, respectively. The two latent codes produced by  $E_{LR}$  and  $E_{HR}$  are combined in a style transfer way (AdaIN) rather than concatenated directly. The style information  $y$  (i.e.,  $Z_{LR}$ ) controls AdaIN [7] operations after each residual block in the generator  $G$ .

Following VAE, we use the Kullback-Leibler (KL) divergence to regularize the latent space obtained by  $E_{LR}$ . The  $E_{LR}$  branch has two output variables, i.e.,  $\mu$  and  $\sigma$ . To a reparameterization trick, we have  $Z_{LR} = \mu_{LR} + \sigma \odot \epsilon$ , where  $\epsilon \sim \mathcal{N}(0, I)$ ,  $\odot$  means Hadamard product;  $\mu$  and  $\sigma$  denote the mean and the standard deviation, respectively. Given  $N$  data samples, the posterior distribution  $q_\phi(z|x_{LR})$  is constrained through Kullback-Leibler divergence:

$$\mathcal{L}_{kl} = KL(q_\phi(z|x_{LR})||p(z)) = \frac{1}{2N} \sum_{i=1}^N \sum_{j=1}^M (1 + \log(\sigma_{ij}^2) - \mu_{ji}^2 - \sigma_{ij}^2), \quad (1)$$



**Fig. 2.** The proposed GA module in GARB. The  $\otimes$  denotes the matrix multiplication operation.

where  $q_\phi(\cdot)$  is the inference network  $E_{LR}$ . The prior  $p(z)$  is the standard multi-variate Gaussian distribution.  $M$  is the dimension of  $z$ .

The generator  $p_\theta(y_{LR}|z_{LR}, z_{HR})$  in styleVAE is required to generate LR images  $y_{LR}$  from the latent space  $z_{HR}$  and the learned distribution  $z_{LR}$ . Similar [13, 14], we use a pre-trained VGG network [23] to calculate the following loss function:

$$\mathcal{L}_{style} = \alpha \mathcal{L}_c + \beta \mathcal{L}_s, \quad (2)$$

where  $\alpha$  and  $\beta$  are the weights for the content loss and the style loss, respectively. Here we set  $\alpha$  and  $\beta$  to 1 and 0.1, respectively. It defines at a specific layer  $J$  of the VGG network [23]:

$$\mathcal{L}_c = \|\phi^J(y_{LR}) - \phi^J(x_{HR})\|_F^2, \quad (3)$$

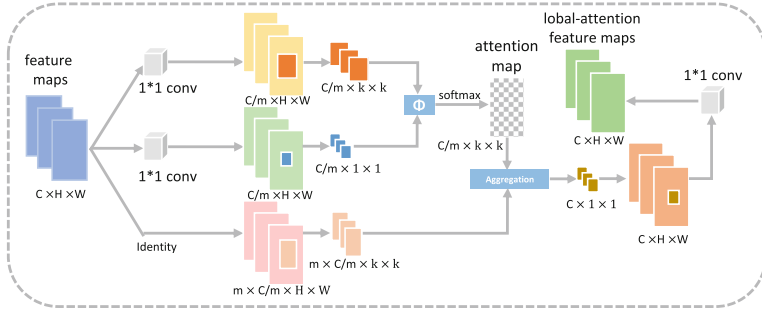
where  $y_{LR}$  and  $x_{HR}$  denote generated the LR images and the corresponding HR images, respectively. We resize the size of  $x_{HR}$  to match that of  $y_{LR}$ . Furthermore,  $\mathcal{L}_s$  is defined by a weighted sum of the style loss at different layers of the pre-trained VGG network:

$$\mathcal{L}_s = \sum_i w_i \mathcal{L}_s^i(y_{LR}, x_{LR}), \quad (4)$$

where  $w_i$  is the trade-off factor for the style loss  $\mathcal{L}_s^i$  at  $i$ th layer of the pre-trained VGG network.  $\mathcal{L}_s^i$  is computed as the Euclidean distance between the Gram matrices of the feature maps for  $y_{LR}$  and  $x_{LR}$ .

**Mutual Information Maximization.** The purpose of the inference network  $E_{LR}$  is to extract the style information. To gain style representation better, the mutual information between real-world and generated LR images is required to be maximized as follows:

$$\mathcal{L}_{mi} = \sup_{\theta \in \Theta} \mathbb{E}_{p(x_{LR}, y_{LR})} [T_\theta] - \log(\mathbb{E}_{p(x_{LR}) \otimes p(x_{HR})} [e^{T_\theta}]), \quad (5)$$



**Fig. 3.** The proposed LA module in LARB. Different from GA, it obtains the composability between a target pixel and a pixel from the visible scope of the target pixel instead of all pixels of images.

where  $T_\theta$  denotes a static deep neural network parameterized by  $\theta \in \Theta$ . The inputs of the  $T_\theta$  are empirically drawn from the joint distribution  $p_{(x_{LR}, y_{LR})}$  and the product of the marginal  $p_{x_{LR}} \otimes p_{y_{LR}}$ .

According to all the loss functions mentioned above, the overall loss to optimize styleVAE is formulated as:

$$\mathcal{L}_{styleVAE} = \lambda_1 \mathcal{L}_{kl} + \lambda_2 \mathcal{L}_{style} + \lambda_3 \mathcal{L}_{mi}, \tag{6}$$

where  $\lambda_1$ ,  $\lambda_2$ , and  $\lambda_3$  are the trade-off factors.

### 3.2 Super-Resolution Network

Our proposed SR network mainly consists of global attention residual block (GARB) and local attention residual block (LARB).

**Global Attention Residual Block.** As shown in Fig. 2, we propose a global attention residual block to learn long-range dependencies by using the global attention (GA) module [25]. It maintains efficiency in calculation and statistics. There is a skip connection in GARB due to the success of residual blocks (RBs) [27] (See Fig. 1). The GA module is formulated as follows:

$$\beta_{j,i} = \frac{\exp(s_{i,j})}{\sum_{i=1}^N \exp(s_{i,j})}, \tag{7}$$

where  $s_{i,j} = f(x_i)^T g(x_j)$  ( $f(x) = W_f x$ ,  $g(x) = W_g x$ ) represents that the feature maps of the former hidden layer are projected into two latent spaces to obtain the attention value.  $\beta_{i,j}$  indicates the degree of attention that the  $i^{th}$  position receives when generating the  $j^{th}$  area. The output of the attention layer is defined as:

$$o_j = v \left( \sum_{i=1}^N \beta_{(j,i)} h(x_i) \right), \tag{8}$$

where  $h(x_i) = W_h x_i$ ,  $v(x_i) = W_v x_i$ . The above  $W_f$ ,  $W_g$ ,  $W_h$ ,  $W_v$  are implemented by a convolution layer with kernel size  $1 \times 1$ . We connect  $o_i$  and  $x_i$  in a residual way, so the final output is shown as below:

$$y_i = \lambda o_i + x_i, \quad (9)$$

where  $\lambda$  is a learnable scalar.

**Local Attention Residual Block.** As shown in Fig. 3, we also propose a local attention residual block (LARB) to capture local details through the local attention (LA) module [9]. The LA module forms local pixel pairs with a flexible bottom-up way, which efficiently deals with visual patterns with increasing size and complexity. We use a general method of relational modeling to calculate the LA module, which is defined as:

$$\omega(p', p) = \text{softmax}(\Phi(f_{\theta_q}(x_{p'}), f_{\theta_k}(x_p))), \quad (10)$$

where  $\omega(p', p)$  obtains a representation at one pixel by computing the composability between it (target pixel  $p'$ ) and a pixel  $p$  in its visible position range. Transformation functions  $f_{\theta_q}$  and  $f_{\theta_k}$  are implemented by  $1 \times 1$  convolution layer. The function  $\Phi$  is chosen the squared difference:

$$\Phi(q_{p'}, k_p) = -(q_{p'} - k_p)^2. \quad (11)$$

## 4 Experiments

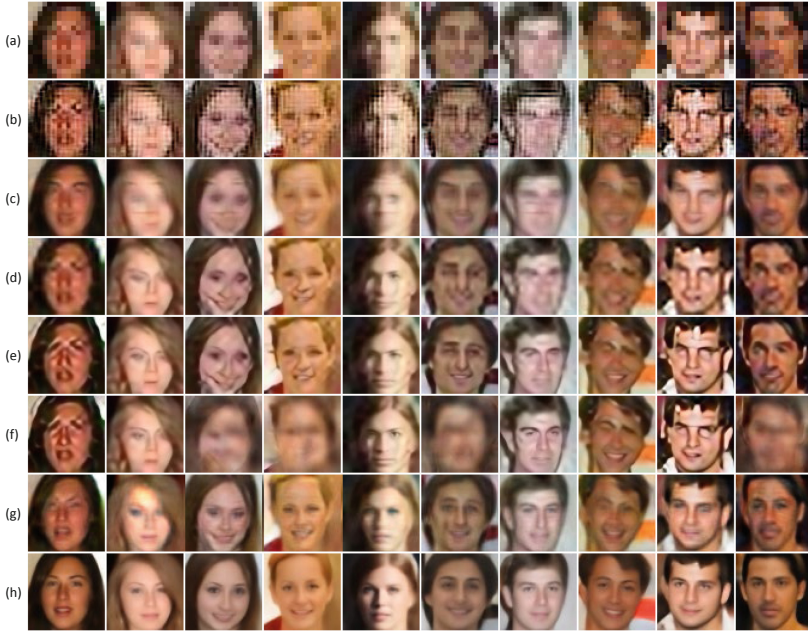
In this section, we firstly introduce the datasets and implementation in detail. Then we evaluate our proposed method from both qualitative and quantitative aspects.

### 4.1 Datasets and Implementation

**Training Dataset.** As illustrated in [3], we select the following four datasets to build the HR training dataset that contains 180k faces. The first is a subset of VGGFace2 [4] that contains images with 10 large poses for each identity (9k identities). The second is a subset of Celeb-A [22] that contains 60k faces. The third is the whole AFLW [17] that contains 25k faces used for facial landmark localization originally. The last is a subset of LS3D-W [1] that contains face images with various poses.

We also utilize the WIDER FACE [24] to build a real-world LR dataset. WIDER FACE is a face detection benchmark dataset that consists of more than 32k images affected by various noise and degradation types. We randomly select 90% images in the LR training dataset.

**Testing Dataset.** Another 10% images from WIDER FACE described in the latest section are selected as real-world LR testing dataset. We conduct experiments on it to verify the performance of our proposed method.



**Fig. 4.** Comparisons with the state-of-art methods ( $4\times$ ). (a) Input real-world LR images. (b) Results of SRCNN [6]. (c) Results of SRGAN [20]. (d) Results of VDSR [15]. (e) Results of EDSR [21]. (f) Results of RDN [27]. (g) Results of [3]’s method (h) Our results. Compared with other methods, our proposed pipeline reconstructs sharper SR images with better details.

**Implementation Details.** Our proposed styleVAE is trained on the unpaired training HR and LR images for 10 epochs. After that, the paired LR-HR dataset is created used to train the SR network with 50 epochs. We train our styleVAE and SR network through the ADAM algorithm [16] with  $\beta_1 = 0.9$ ,  $\beta_2 = 0.999$ . The learning rate is initially set to  $10^{-4}$  and remains unchanged during the training process. The weights  $\lambda_1$ ,  $\lambda_2$ , and  $\lambda_3$  are set as 0.01, 1, and 0.1, respectively. We use PyTorch to implement our models and train them on NVIDIA Titan Xp GPUs.

## 4.2 Experimental Results

**Real-World Images.** In this section, We conduct experiments on the testing dataset described in Sect. 4.1. In order to evaluate the performance of our proposed method, we compare with the following state-of-the-art methods both numerically and qualitatively: SRCNN [6], SRGAN [20], VDSR [15], EDSR [21], Bulat’s method [3] and RDN [26]. We retrain all these compared methods for the sake of fairness on our HR training dataset with the default configurations described in their respective papers. Note that LR images are produced by applying a bicubic kernel to the corresponding HR images.

**Table 1.** Results of different SR methods. The second and third columns show PSNR and SSIM based performance on synthetic real-world-like LR images (Higher is better). The fourth column shows FID based performance on our testing dataset (Lower is better).

Method	PSNR	SSIM	FID
BICUBIC	21.19	0.5570	288.47
SRCNN [6]	19.57	0.4030	256.78
SRGAN [20]	20.36	0.5007	179.70
VDSR [15]	20.61	0.5584	144.29
EDSR [21]	20.44	0.5137	129.93
RDN [27]	18.18	0.4063	162.04
Bulat’s [3]	22.76	0.6296	149.97
Ours	<b>24.16</b>	<b>0.7197</b>	<b>98.61</b>

**Table 2.** Results of experiments on deep plug-and-play SR in FID. (Lower is better).

Data Type	Scale	SRCNN	EDSR
BICUBIC	$\times 4$	256.78	129.93
styleVAE	$\times 4$	198.75	107.63

In numerical terms, we use Fréchet Inception Distance (FID) [8] to measure the quality of the generated images since there are no corresponding HR images. The quantitative results of different SR methods on our testing dataset are summarized in Table 1 (with the factor  $\times 4$ ). It clearly demonstrates that our proposed method is superior to other compared approaches and achieves the best performance on the testing dataset. We also discover that the performances of compared methods trained on bicubic-downsampled LR images are degraded when applied to real-world LR images. The main reason is that nuisance factors, e.g. motion blur, lens aberration and sensor noise, are not taken into synthetic LR images by bicubic interpolation. By training on real-world-like LR images, our method is superior to all compared methods, and the FID value is reduced by a maximum of 158.17.

In Fig. 4, we visually show the qualitative results the our testing dataset with  $\times 4$  scale. There are significant artifacts in HR images generated by shallower networks, e.g. SRCNN [6] and SRGAN [20]. Serious mesh phenomenons are found in reconstructed images by SRCNN. We also discover that generated images of VDSR [15], EDSR [21] and RDN [27] are usually distorted. On the contrary, SR images generated by our proposed method are more realistic than Bulat’s method [3], since LR images produced by styleVAE exceedingly resemble real-world LR images.

**Real-World-Like Images.** In order to verify the performance of the proposed method on LR images with unknown degradation modes, we conduct exper-



iment on synthetic real-world-like LR images obtained by styleVAE with  $\times 4$  scale. We utilize images from LFW [10–12, 19] as the HR image inputs of styleVAE to generate real-world-like LR images. The second and third columns of Table 1 report PSNR and SSIM results of different SR methods. We find that the performances of compared methods are very limited, even lower than that of bicubic up-sampling directly. It also demonstrates that simulating real-world-like LR images is an effective way to improve performance when applied to real-world LR images.

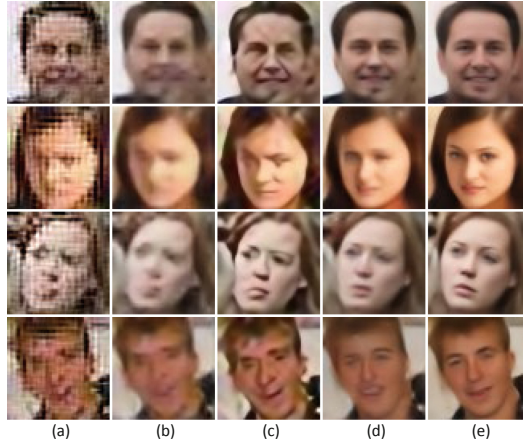
**Deep Plug-and-Play Super-Resolution.** To further validate the effectiveness of styleVAE, we design two pipelines with the help of plug-and-play framework. We can simply plug styleVAE into SR networks to replace bicubic down-sampled LR images that are used in many previous SR methods. We choose two of the compared methods as the plugged SR networks: a shallower SR network SRCNN [6] and a deeper SR network EDSR. Thus there are four versions of SR networks: SRCNN-B and EDSR-B, trained on bicubic down-sampled LR images, SRCNN-S and EDSR-S, trained on LR images generated by styleVAE. The FID results on our testing dataset are reported in Table 2. As shown in Table 2, the FID values of SRCNN-B and EDSR-B (the second row of Table 2) are higher than those of EDSR-S and SRCNN-S (the last row of the Table 2). By simulating real-world LR images using styleVAE, SRCNN gains an improvement of 58.3 (the third column of Table 2) and EDSR is improved by 22.3 (the last column of Table 2).

We also demonstrate the visual results in Fig. 5. As shown in Fig. 5, compared (a) with (b), SRCNN-S effectively eliminates the mesh phenomenon in the image generated by SRCNN-B. When training on LR images generated by styleVAE, EDSR-S produces more pleasing results (d) rather than distorted reconstructed images (c) by EDSR-B. Compared (b), (d) and (e), our proposed method is able to generate sharper images.

### 4.3 Ablation Studies

**StyleVAE.** In order to investigate the effectiveness of the mutual information estimation (MI) and Adaptive Instance Normalization (AdaIN) used in styleVAE, we train several other variants of styleVAE: remove MI or/and AdaIN. To evaluate the performance of these variants of styleVAE, we measure FID between LR images generated by these variants and real-world LR images from our testing dataset. FID results are provided in Table 3. When both AdaIN and MI are removed, the FID value is relatively high. After arbitrarily adding one of the two, the value of FID is decreased. For both MI and AdaIN used in styleVAE, the FID result is the lowest. We also evaluate how similar the synthetic LR images by bicubic down-sample and real-world LR images from WIDER FACE. The FID result is found as 31.20. These results faithfully indicate that AdaIN and MI are essential for styleVAE.

**SR Network.** Similar to the ablation investigation of styleVAE, we also train several variants of the proposed SR network: remove the GA or/and LA module(s) in the SR network. These several variants are trained on LR images



**Fig. 5.** Results of experiments on deep plug-and-play super-resolution. (a) Results of SRCNN-B. (b) Results of SRCNN-S. (c) Results of EDSR-B. (d) Results of EDSR-S. (e) Results of our method. This suggests that the performance of SR methods can be improved by training on LR images generated by styleVAE.

**Table 3.** Investigations of AdaIN and MI in styleVAE. We also evaluate how similar the synthetic LR images by bicubic down-sample and real-world LR images in WIDER FACE. The FID value between these is 31.20 (Lower is better). Results of experiments on deep plug-and-play SR in FID. (Lower is better).

AdaIN	×	×	✓	✓
MI	×	✓	×	✓
FID	26.6	25.78	24.63	18.77

produced by performing bicubic interpolation on corresponding HR images. In Table 4, when both the GA and LA modules are removed, the PSNR value on LFW (with upscale factor  $\times 4$ ) is the lowest. When the LA module is added, the PSNR value is increased by 0.1 dB. After adding the GA module, the performance reaches 30.27 dB. When both attention modules are added, the performance is the best, with a PSNR of 30.43 dB. These experimental results clearly demonstrate that these two attention modules are necessary for the proposed SR network and greatly improve its performance (Table 4).

**Table 4.** Ablation investigation on the effects of the GA and LA modules in SR network. The PSNR (dB) values are reported on LFW (higher is better).

GA	×	×	✓	✓
LA	×	✓	×	✓
PSNR	30.08	30.18	30.27	30.43

## 5 Conclusion

We have proposed a novel two-stage process to address the challenging problem of super-resolving real-world LR images. The proposed pipeline unifies a style-based Variational Autoencoder (styleVAE) and an SR network. Due to the participation of nuisance factors transfer and VAE, styleVAE generates real-world-like LR images. Then the generated LR images paired with the corresponding HR images are fed into the SR network. Our SR network firstly learns long-range dependencies by using GARB. Then the attention of SR network moves to local areas of images in which texture detail will be filled out through LARB. Extensive experiments show our superiority over existing state-of-the-art SR methods and the ability of styleVAE to facilitate method generalization and robustness to real-world cases.

## References

1. Bulat, A., Tzimiropoulos, G.: How far are we from solving the 2D & 3D face alignment problem? (and a dataset of 230,000 3D facial landmarks). In: International Conference on Computer Vision (2017)
2. Bulat, A., Tzimiropoulos, G.: Super-FAN: integrated facial landmark localization and super-resolution of real-world low resolution faces in arbitrary poses with GANs. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2018)
3. Bulat, A., Yang, J., Tzimiropoulos, G.: To learn image super-resolution, use a GAN to learn how to do image degradation first. In: Proceedings of the European Conference on Computer Vision (ECCV) (2018)
4. Cao, Q., Shen, L., Xie, W., Parkhi, O.M., Zisserman, A.: VGGFace2: a dataset for recognising faces across pose and age. In: IEEE International Conference on Automatic Face & Gesture Recognition (FG) (2018)
5. Chen, Y., Tai, Y., Liu, X., Shen, C., Yang, J.: FSRNet: end-to-end learning face super-resolution with facial priors. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2492–2501 (2018)
6. Dong, C., Loy, C.C., He, K., Tang, X.: Image super-resolution using deep convolutional networks. *IEEE Trans. Patt. Anal. Mach. Intell.* **38**, 295–307 (TPAMI) (2015)
7. Gatys, L.A., Ecker, A.S., Bethge, M.: Image style transfer using convolutional neural networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2414–2423 (2016)
8. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.: GANs trained by a two time-scale update rule converge to a local Nash equilibrium. In: Conference on Neural Information Processing Systems (NeurIPS) (2017)
9. Hu, H., Zhang, Z., Xie, Z., Lin, S.: Local relation networks for image recognition. arXiv preprint [arXiv:1904.11491](https://arxiv.org/abs/1904.11491) (2019)
10. Huang, G.B., Jain, V., Learned-Miller, E.: Unsupervised joint alignment of complex images. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV) (2007)
11. Huang, G.B., Mattar, M., Lee, H., Learned-Miller, E.: Learning to align from scratch. In: Conference on Neural Information Processing Systems (NeurIPS) (2012)

12. Huang, G.B., Ramesh, M., Berg, T., Learned-Miller, E.: Labeled faces in the wild: a database for studying face recognition in unconstrained environments. University of Massachusetts, Amherst, Technical report (2007)
13. Huang, H., He, R., Sun, Z., Tan, T.: Wavelet domain generative adversarial network for multi-scale face hallucination. *Int. J. Comput. Vis.* **127**(6–7), 763–784 (2019)
14. Huang, X., Belongie, S.: Arbitrary style transfer in real-time with adaptive instance normalization. In: *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pp. 1501–1510 (2017)
15. Kim, J., Lee, J.K., Mu Lee, K.: Accurate image super-resolution using very deep convolutional networks. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1646–1654 (2016)
16. Kingma, D.P., Ba, J.: Adam: a method for stochastic optimization. arXiv preprint [arXiv:1412.6980](https://arxiv.org/abs/1412.6980) (2014)
17. Koestinger, M., Wohlhart, P., Roth, P.M., Bischof, H.: Annotated facial landmarks in the wild: a large-scale, real-world database for facial landmark localization. In: *IEEE International Conference on Computer Vision Workshops (ICCV workshops)* (2011)
18. Kotovenko, D., Sanakoyeu, A., Lang, S., Ommer, B.: Content and style disentanglement for artistic style transfer. In: *Proceedings of the IEEE International Conference on Computer Vision (ICCV)* (2019)
19. Learned-Miller, G.B.H.E.: Labeled faces in the wild: Updates and new reporting procedures. University of Massachusetts, Amherst, Technical report (2014)
20. Ledig, C., 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 (CVPR)*, pp. 4681–4690 (2017)
21. Lim, B., Son, S., Kim, H., Nah, S., Lee, K.M.: Enhanced deep residual networks for single image super-resolution. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, July 2017
22. Liu, Z., Luo, P., Wang, X., Tang, X.: Deep learning face attributes in the wild. In: *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pp. 3730–3738 (2015)
23. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint [arXiv:1409.1556](https://arxiv.org/abs/1409.1556) (2014)
24. Yang, S., Luo, P., Loy, C.C., Tang, X.: Wider face: a face detection benchmark. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5525–5533 (2016)
25. Zhang, H., Goodfellow, I., Metaxas, D., Odena, A.: Self-attention generative adversarial networks. arXiv preprint [arXiv:1805.08318](https://arxiv.org/abs/1805.08318) (2018)
26. Zhang, Y., Li, K., Li, K., Wang, L., Zhong, B., Fu, Y.: Image super-resolution using very deep residual channel attention networks. In: *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 286–301 (2018)
27. Zhang, Y., Tian, Y., Kong, Y., Zhong, B., Fu, Y.: Residual dense network for image super-resolution. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2472–2481 (2018)
28. Zhu, S., Liu, S., Loy, C.C., Tang, X.: Deep cascaded Bi-network for face hallucination. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) *ECCV 2016*. LNCS, vol. 9909, pp. 614–630. Springer, Cham (2016). [https://doi.org/10.1007/978-3-319-46454-1\\_37](https://doi.org/10.1007/978-3-319-46454-1_37)