Supplementary Material
- VarSR: Variational Super-Resolution Network for Very Low Resolution Images

Sangeek Hyun[1] and Jae-Pil Heo[1,2]

[1] Dept. of Artificial Intelligence, Sungkyunkwan University
[2] Dept. of Computer Science and Engineering, Sungkyunkwan University

1 Cropped CelebA

Since the background of facial image is irrelevant to the semantic of the human face, images having fairly large background regions are barely suitable for validation of super-resolution performance. Thus, those background regions can dilute the benefit of our method which is capable of generative diverse results to overcome the underlying uncertainty of low resolution images. Therefore, we perform extra cropping to CelebA dataset to focus solely on the human face rather than backgrounds. Specifically, the raw images of 64×64 pixels are cropped 10 pixels for all directions so that their resolutions become 44×44 pixels. We denote the altered dataset as Cropped CelebA.

As reported in Table. 1, our method significantly outperforms the baseline techniques in all the tested metrics. Interestingly, our method shows the best performance on traditional metrics (PSNR, SSIM, and MSE) even with the adversarial loss. Those experimental results confirm that our method can super-resolve very low resolution facial images more accurately if the images are solely focused on semantic parts of human face without backgrounds. Note that, the model architectures and evaluation protocols utilized for this experiments are exactly identical to ones for Table. 1 of the main manuscript.

Table 1. Quantitative results on Cropped CelebA dataset with adversarial loss. Cropped CelebA dataset has images of 44×44 pixels center-cropped from original images. For Ours and MR-GAN, the PSNR, SSIM, and MSE scores are computed with the image having the best PSNR among 10 samples, while the best LPIPS and Facenet scores within 10 samples are reported.

	SRGAN	MR-GAN	Ours (λ_{KL}=1e-2)	Ours (λ_{KL}=2e-2)
PSNR	23.08	22.28	23.15	**23.79**
SSIM	0.7490	0.7050	0.7503	**0.7788**
MSE	74.21	76.89	72.85	**71.19**
LPIPS	0.0433	0.0310	**0.0283**	0.0291
Facenet	0.0469	0.0436	**0.0421**	0.0428

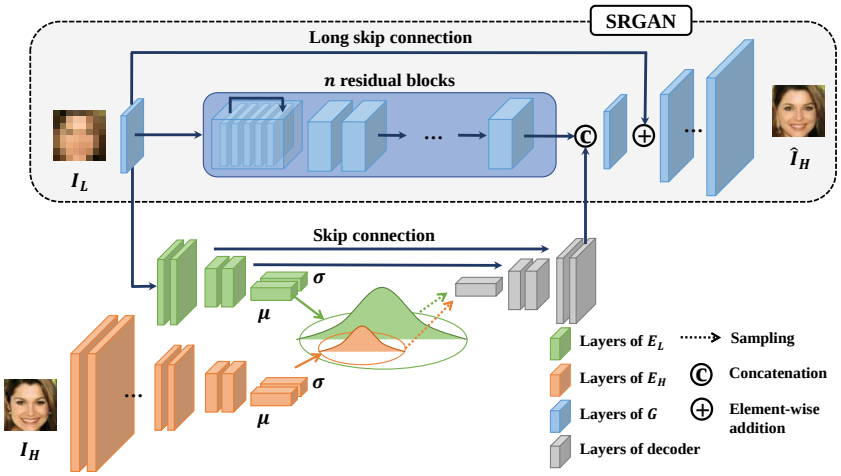

Fig. 1. The architecture of VarSR-Net for the face super-resolution.

2 Model Architectures

Our generator model G has the same network architecture and capacity with the SRGAN except it additionally receives a sampled latent representation.

Our model has two encoders that predict the parameters of multivariate Gaussian distribution, E_L and E_H, from low and high resolution images, respectively. A common decoder produces a tensor that has the same dimension with the output of the final residual block of G from a sampled latent vector. The tensor computed by the decoder is, then, passed to the G by concatenation to the output of the last residual block of G. The whole network architecture for the face super-resolution is illustrated in Fig. 1.

In detail, our encoder and decoder have two consecutive convolution layers for each spatial level for the face dataset. Pooling and upsampling are performed by strided and subpixel convolution operations, respectively. The encoder has two additional fully-connected layers to predict the parameters of multivariate Gaussian distribution.

In digits datasets, we utilize only one convolution layer at each spatial level. We also remove the residual blocks and a long skip connection from the generator for all the tested methods including ours to adapt to extremely low dimensionality of digit images.

3 Additional Results

We report additional qualitative results in Fig. 2 (on CelebA with λ_{KL}=1e-2), Fig. 3 (on CelebA with λ_{KL}=2e-2), Fig. 4 (on MNIST), and Fig. 5 (on LP). Specifically, we show 5 and 10 randomly sampled results in digits and human face benchmarks, respectively.

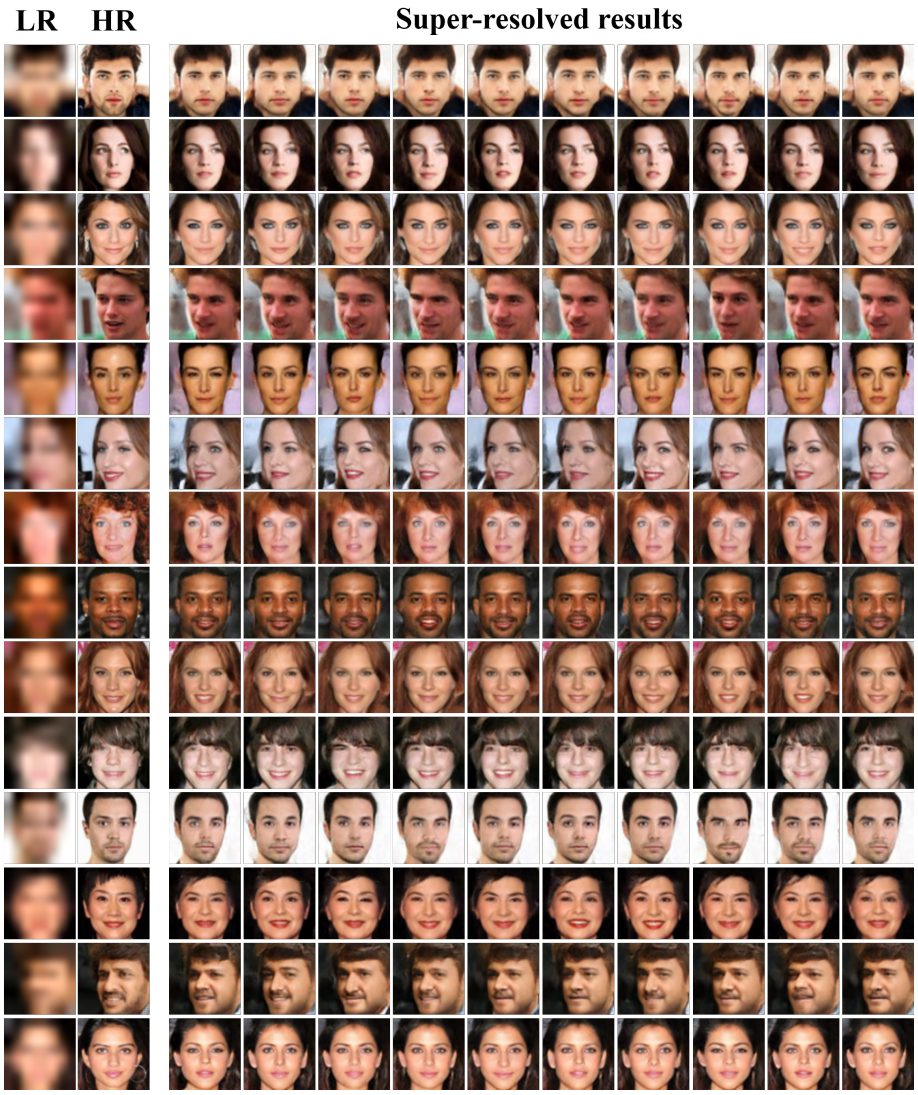

Fig. 2. Qualitative results on CelebA dataset with λ_{KL}=1e-2. Zoom in for details.

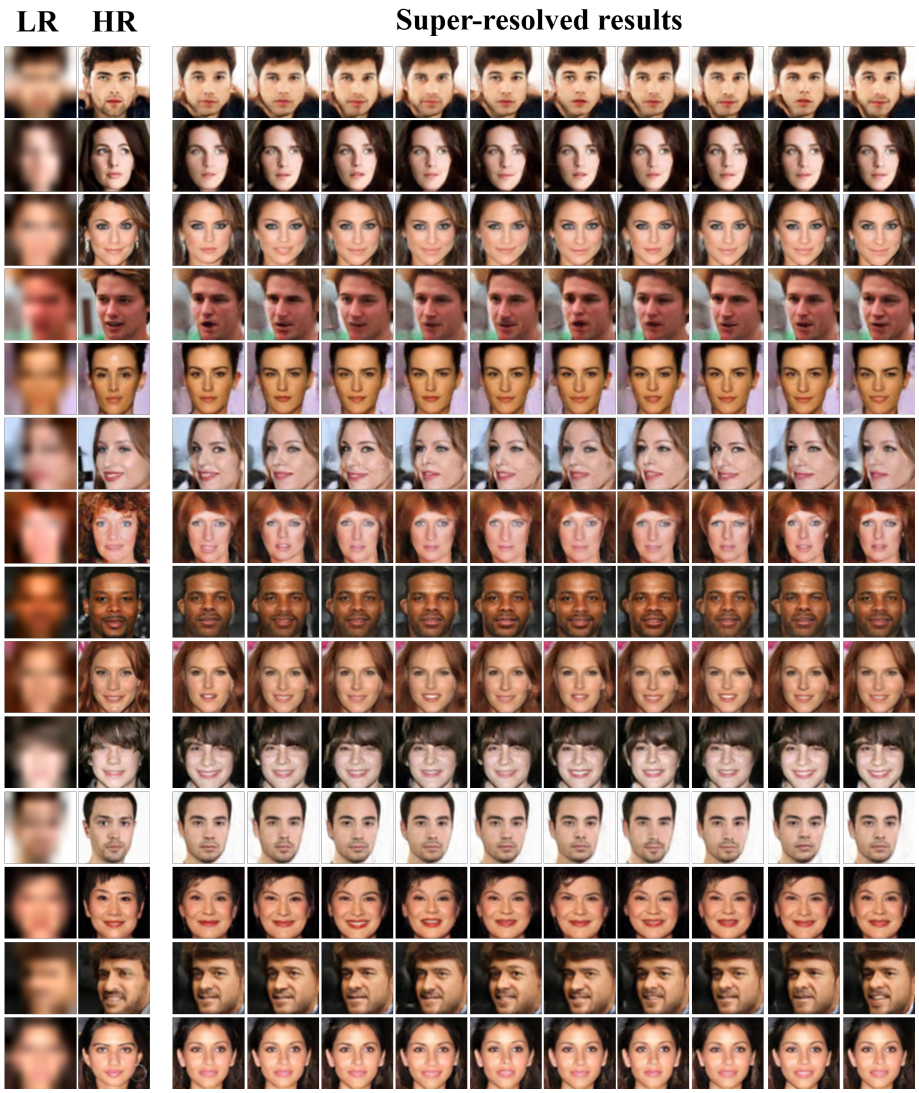

Fig. 3. Qualitative results on CelebA dataset with λ_{KL}=2e-2. Zoom in for details.

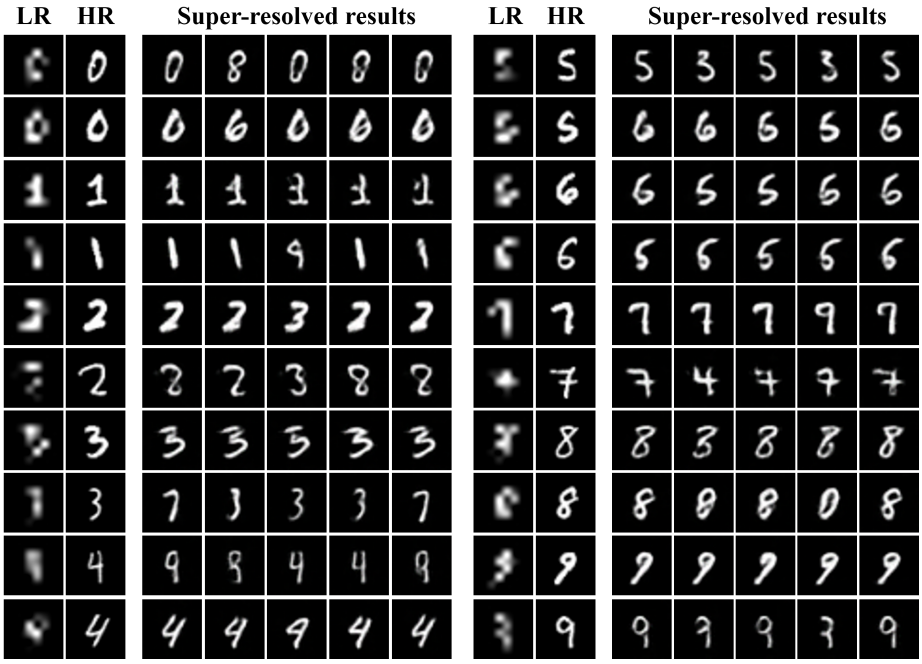

Fig. 4. Qualitative results on MNIST dataset. Zoom in for details.

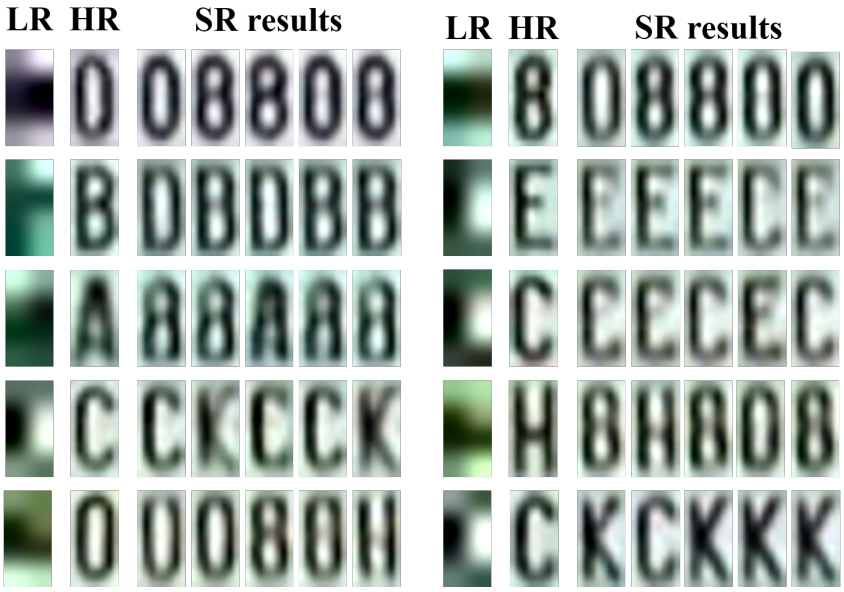

Fig. 5. Qualitative results on LP dataset. Zoom in for details.