Multi-Scale Face Restoration with Sequential Gating Ensemble Network

Jianxin Lin, Tiankuang Zhou, Zhibo Chen

University of Science and Technology of China, Hefei, China

{linjx, zhoutk}@mail.ustc.edu.cn, chenzhibo@ustc.edu.cn

Abstract

Restoring face images from distortions is important in face recognition applications and is challenged by multiple scale issues, which is still not well-solved in research area. In this paper, we present a Sequential Gating Ensemble Network (SGEN) for multi-scale face restoration issue. We first employ the principle of ensemble learning into SGEN architecture design to reinforce predictive performance of the network. The SGEN aggregates multi-level base-encoders and base-decoders into the network, which enables the network to contain multiple scales of receptive field. Instead of combining these base-en/decoders directly with non-sequential operations, the SGEN takes base-en/decoders from different levels as sequential data. Specifically, the SGEN learns to sequentially extract high level information from base-encoders in bottom-up manner and restore low level information from base-decoders in top-down manner. Besides, we propose to realize bottom-up and top-down information combination and selection with Sequential Gating Unit (SGU). The SGU sequentially takes two inputs from different levels and decides the output based on one active input. Experiment results demonstrate that our SGEN is more effective at multi-scale human face restoration with more image details and less noise than state-of-the-art image restoration models. By using adversarial training, SGEN also produces more visually preferred results than other models through subjective evaluation.

1 Introduction

In the past decades, facial analysis techniques, such as face recognition and face detection, have achieved great progress. Meanwhile, thanks to the rapid development of surveillance system, facial analysis techniques have been employed to various applications, such as criminal investigation. However, the performance of most facial analysis techniques may degrade rapidly when given low quality face images. In real surveillance systems, the quality of the surveillance face images are affected by many factors including low resolution, blur and noise. Therefore, how to restore a high quality face from a low quality face is a challenge. Face restoration technique provides a viable way to improve performance of facial analysis techniques on low quality face images.
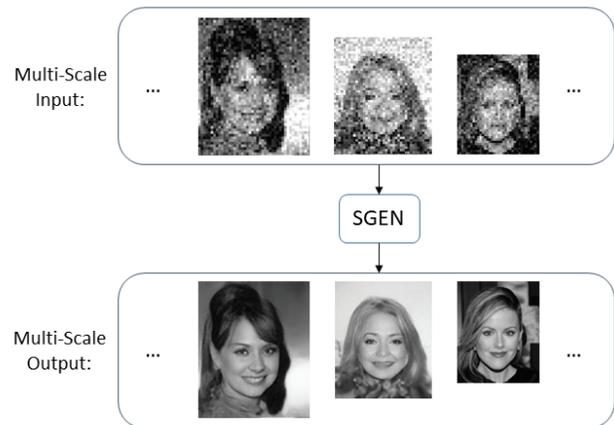

Figure 1: Illustration of our SGEN on multi-scale face restoration. The multi-scale noise corrupted LR face images are up-sampled to the size of ground truth before being fed into network.

Since face restoration has great potential in real applications, numerous face restoration algorithms have been proposed in recent years. Some algorithms focus on solving face restoration from low-resolution (LR) problem, such as works in (Wang and Tang 2005; Ma, Zhang, and Qi 2010; Wang et al. 2014; Yu and Porikli 2016). Other algorithms also take the noise corruption into consideration during face super resolution, such as works in (Jiang et al. 2014; 2016). We observe that most existing face restoration methods omit one vital characteristic of real-world images, namely images in real applications always contain faces of different scales. Also, when the images are corrupted with serious distortions, it's hard to extract the faces in the distorted images for face restoration since face detection methods may not work well under this situation. Therefore, in this paper, we focus on solving face restoration close to real-world situation as illustrated in Figure 1. Our proposed model can effectively restore face images with details from noise corrupted LR face images without scale limitation.

Face restoration can be considered as an image-to-image translation problem that transfers one image domain to an-

other image domain. Solutions on image-to-image translation problem (Taigman, Polyak, and Wolf 2016; Yoo et al. 2016; Johnson, Alahi, and Fei-Fei 2016) usually use autoencoder network (Hinton and Salakhutdinov 2006) as a generator. However, single autoencoder network is too simple to represent multi-scale image-to-image translation due to lack of multi-scale representation. Meanwhile, ensemble learning, a machine learning paradigm where multiple learners are trained to solve the same problem, have shown its ability to make accurate prediction from multiple "weak learners" in classification problem (Dietterich 2000; Kuncheva 2004; Polikar 2006). Therefore, an effective way to reinforce predictive performance of autoencoder network can be aggregating multiple base-generators into an enhanced-generator. In our model, we introduce base-encoders and base-decoders from low level to high level. These multi-level base-en/decoders ensure the generator have more diverse representation capacity to deal with multi-scale face image restoration.

The typical way of ensemble is to take a vote for the outputs of base-en/decoders. However, multi-scale face restoration is a problem that concerns multiple processes of feature extracting and restoring, merely taking a vote (or with other ensemble method) fails to incorporate high level information and restore detail information. Based on this observation, we devise a sequential ensemble structure that takes base-en/decoders from different levels as sequential data. The different combination directions of base-en/decoders are determined by the different goals of encoder and decoder. This sequential ensemble method is inspired by long short-term memory (LSTM) (Hochreiter and Schmidhuber 1997). LSTM has been proved successfully in modeling sequential data, such as text and speech (Sundermeyer, Schlüter, and Ney 2012; Sutskever, Vinyals, and Le 2014). The LSTM has the ability to optionally choose information passing through because of the gate mechanism. Specially, we design a Sequential Gating Unit (SGU) to realize information combination and selection, which sequentially takes base-en/decoders' information from two different levels as inputs and decides the output based on one active input.

Traditional optimization target of image restoration problem is to minimize the mean square error (MSE) between the restored image and the ground truth. However, minimizing MSE will often encourage smoothness in the restored image. Recently, generative adversarial networks (GANs) (Goodfellow et al. 2014; Denton et al. 2015; Radford, Metz, and Chintala 2015; Salimans et al. 2016) show state-of-the-art performance on generating pictures of both high resolution and good semantic meaning. The high level real/fake decision made by discriminator causes the generated images looking real to the class of target domain. Therefore, we utilize the adversarial learning process proposed in GAN (Goodfellow et al. 2014) for restoration model training.

In general, we propose to solve multi-scale face restoration problem with a Sequential Gating Ensemble Network (SGEN). The contribution of our approach includes three aspects:

- We employ the principle of ensemble learning into network architecture design. The SGEN is composed of multi-level base-en/decoders, which has better representation ability than ordinary autoencoder.

- The SGEN takes base-en/decoders from different levels as sequential data with two different directions corresponding to the different goals of encoder and decoder, which enables network to learn more compact high level information and restore more low level details.

- Furthermore, we propose a SGU unit to sequentially guide the combination of information from different levels.

The rest of this paper is organized as follows. We introduce related work in Section 2 and present the details of proposed SGEN in Section 3, including network architecture, SGU unit and adversarial learning for SGEN. We present experiment results in Section 4 and conclude in Section 5.

2 Related Work

Face restoration has been studied for many years. In the early time, (Wang and Tang 2005) utilized global face-based method, i.e. principal component analysis (PCA), for face super-resolution. The work in (Ma, Zhang, and Qi 2010) proposed a least squares representation (LSR) framework that restores images using all the training patches, which incorporates more face priors. Due to the instability of LSR, (Wang et al. 2014) introduced a weighted sparse representation (SR) with sparsity constraint for face super-resolution. However, one main drawback of SR based methods is its sensitivity to noise. Accordingly, (Jiang et al. 2014; 2016) introduced to reconstruct noise corrupted LR images with weighted local patch, namely locality-constrained representation (LcR).

Recently, convolutional neural networks (CNNs) based approaches have shown their superior performance in image restoration tasks. SRCNN (Dong et al. 2014) is a three layer fully convolutional network and trained end-to-end for image super resolution. (Yu and Porikli 2016) presented a ultra-resolution discriminative generative network (UR-DGN) that can ultra-resolve a very low resolution face. Instead of building network as simple hierarchy structure, other works also applied the skip connections, which can be viewed as one kind of ensemble structure (Veit, Wilber, and Belongie 2016), to image restoration tasks. (Ledig et al. 2016) proposed a SRResNet that uses ResNet blocks in the generative model and achieves state-of-the-art peak signal-to-noise ratio (PSNR) performance for image super-resolution. In addition, they presented a SRGAN that utilizes adversarial training to achieve better visual quality than SR-ResNet. (Mao, Shen, and Yang 2016) proposed a residual encoder-decoder network (RED-Net) which symmetrically links convolutional and deconvolutional layers with skip-layer connections.

However, these skip-connections in (Ledig et al. 2016; Mao, Shen, and Yang 2016) fail to explore the underlying sequential relationship among multi-level feature maps in image restoration problem. Therefore, we design our SGEN followed by the goal of autoencoder, which sequentially extracts high level information from base-encoders in bottom-up manner and restores low level information from base-decoders in top-down manner.
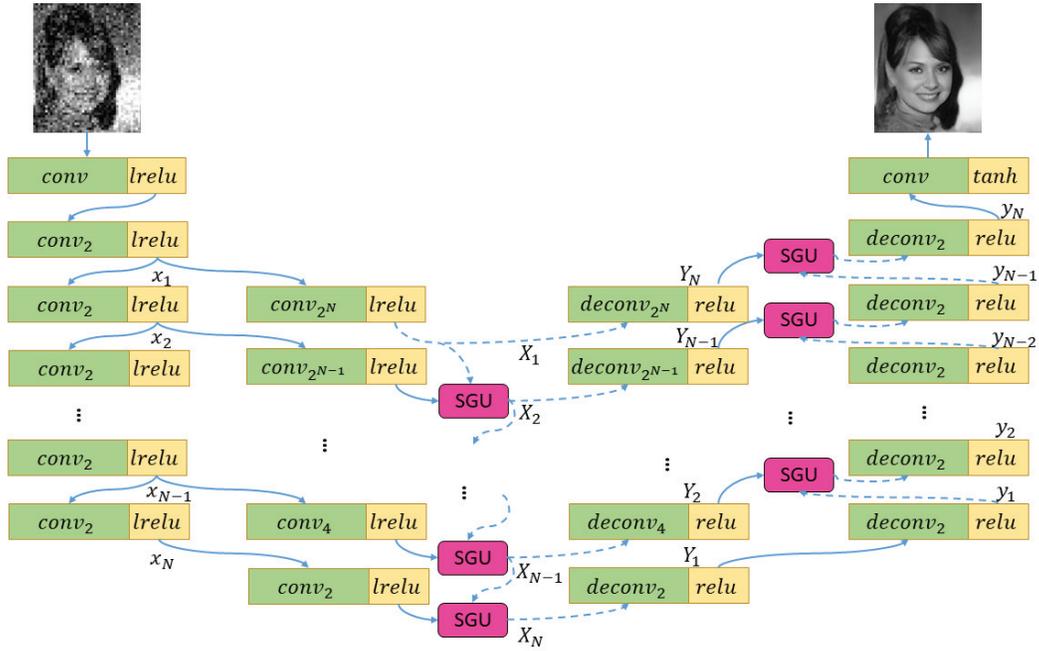
Figure 2: Sequential ensemble network architecture of SGEN. Convolution and pooling operations are shown in green, activation functions are shown in yellow and the SGU is shown in pink.

3 Sequential Gating Ensemble Network

Architecture of our Sequential Gating Ensemble Network (SGEN) is shown in Figure 2. Details are discussed in the following subsections. We first introduce the sequential ensemble network architecture of SGEN. Then we present the Sequential Gating Unit (SGU) combining the multi-level information. Finally, we elaborate the adversarial training for SGEN and the loss function for adversarial training process.

3.1 Sequential ensemble network architecture

First, our generator is a fully convolutional computation network (Long, Shelhamer, and Darrell 2015) that can take arbitrary-size inputs and predict dense outputs. Then, let us denote k-th level encoder feature, k-th level decoder feature, k-th base-encoder feature, k-th base-decoder feature, by x_k, y_k, X_k, Y_k respectively, and suppose there are N base-encoders and base-decoders in total. Given a low quality face image sample s, the SGEN G in Figure 2 can be illustrated in the formulas below:

$$x_1 = lrelu(conv_2(lrelu(conv(s)))), \quad (1)$$

$$x_k = lrelu(conv_2(x_{k-1})), \quad k = 2, 3, ..., N \quad (2)$$

$$X_1 = lrelu(conv_{2^N}(x_1)), \quad (3)$$

$$X_k = SGU(lrelu(conv_{2^{N-k+1}}(x_k)), X_{k-1}), \\ k = 2, 3, ..., N \quad (4)$$

$$Y_k = relu(deconv_{2^k}(X_{N-k+1})), \quad k = 1, 2, 3, ..., N \quad (5)$$

$$y_1 = relu(deconv_2(Y_1)), \quad (6)$$

$$y_k = relu(deconv_2(SGU(Y_k, y_{k-1}))), \quad k = 2, 3, ..., N \quad (7)$$

$$G(s) = tanh(conv(y_N)), \quad (8)$$

where $G(s)$ is the generated face image, $conv_{2^k}$ and $deconv_{2^k}$ are convolution and de-convolution operations with factor 2^k pooling and upsampling respectively. SGU is sequential gating unit. Each de-convolution layer is followed by $relu$ (rectified linear unit) (Nair and Hinton 2010), and each convolution layer is followed by $lrelu$ (leaky relu) (Maas, Hannun, and Ng 2013), except for the last layer of generator (using $tanh$ activation function). Note, there is no parameters sharing among different convolution operations, de-convolution operations and SGUs.

The bottom-up base-encoders combination and top-down base-decoders combination are determined by the different goals of encoder and decoder. Given a low quality face image input, the encoder of a autoencoder would like to transfer the input into highly compact representation with semantic meaning (i.e., bottom-up information extraction), and the decoder would like to restore the face image with abundant details (i.e., top-down information restoration). Therefore, without breaking the rules of autoencoder, we combine the multi-level base-en/decoders in two directions. Accordingly, we design a SGU to realize the multi-level information combination in en/decoder stage respectively.

Combination of these multi-level base-en/decoders provides another benefit that network layer of SGEN contains multiple scales of receptive field, which helps the encoder learn features with multi-scale information and helps decoder generate more accurate images from multi-scale features. Experiment results also demonstrate that our network
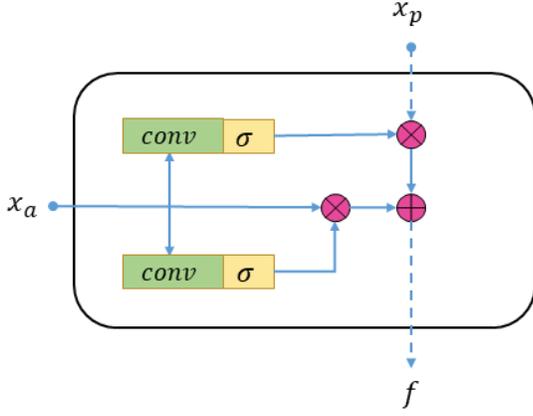
Figure 3: Sequential Gating Unit. Element-wise multiplications and additions are shown in pink.

is more capable of restoring multi-scale low quality face image than other networks.

3.2 Sequential gating unit

To further utilize the sequential relationship among multi-level base-en/decoders, we propose a Sequential Gating Unit (SGU) to sequentially combine and select the multi-level information, which takes base-en/decoders' information from two different levels as inputs and deciding the output based on one active input. The SGU is shown in the Figure 3, equation depicting the unit is given as below:

$$f = \sigma(conv(x_a)) * x_a + \sigma(conv(x_a)) * x_p, \quad (9)$$

where f is the SGU output, $\sigma(x)$ is sigmoid activation function representing "gate" in SGU, x_a and x_p are active input and passive input respectively. The active input x_a makes the decision what information we are going to throw away from the passive input x_p and what new information we are going to add from the active input itself. In the encoder stage, high level base-encoder acts as x_a and takes control over the low level information, which sequentially updates the high level semantic information and removes noise. For the decoder stage, the low level base-decoder becomes x_a and takes control over the high level information in an opposite direction, which sequentially restores low level information and generates images with more details.

3.3 Adversarial training and loss function

We apply the adversarial training of GAN in our proposed model. The adversarial training needs to learn a discriminator D to guide the generator G, i.e. SGEN in this paper, to produce realistic target under the supervision of real/fake. In face restoration case, the objective function of GAN can be represented as minimax function:

$$\min_G \max_D L(D, G) = E_{t \sim p_T(t)}[\log(D(t))] \\ + E_{s \sim p_S(s)}[\log(1 - D(G(s)))], \quad (10)$$

where s is a sample from the low quality source domain S and t is the corresponding sample in high quality target domain T. In addition to using adversarial loss in the generator training process, we add the mean square error (MSE) loss for generator to require the generated image $G(s)$ as close as to the ground truth value of pixels. The modified loss function for adversarial SGEN training is as below:

$$\min_G \max_D L(D, G) = E_{t \sim p_T(t)}[\log(D(t))] \\ + E_{s \sim p_S(s)}[\log(1 - D(G(s)))] \quad (11) \\ + \lambda L_{MSE}(G),$$

$$L_{MSE}(G) = E_{s \sim p_S(s), t \sim p_T(t)}[||t - G(s)||_2^2], \quad (12)$$

where λ is weight to achieve balance between adversarial term and MSE term.

To make the discriminator be able to take input of arbitrary size as well, we design a fully convolutional discriminator with global average pooling proposed in (Lin, Chen, and Yan 2013). We replace the traditional fully connected layer with global average pooling. The idea of global average pooling is to take the average of each feature map as the resulting vector fed into classification layer. Therefore, the discriminator has much fewer network parameters than fully connected network and overfitting is more likely to be avoided.

4 Experiments

4.1 Parameters setting

In the experiments, we set $N = 3$ levels for the SGEN to achieve a trade-off between performance and computation cost and the weight λ is set to 0.1. We use the adaptive learning method Adam (Kingma and Ba 2014) as the optimization algorithm with learning rate of 0.0002. Minibatch size is set to 64 for every experiments.

4.2 Dataset and evaluation metrics

We carry out experiments below on the widely used face dataset CelebA (Liu et al. 2015) containing 202599 cropped celebrity faces. We set aside 30000 images as test set, 20000 images as validation set, and the rest as training set. We resize the face images from 128×96 to 208×176 which are commonly used resolutions in practice. Specially, we sample 6 different scales between two resolutions. Then we downsample the images to low-resolution (LR) by a factor of 4 and corrupt the images by AWGN noise (standard deviation $\sigma = 30$). To enable the noise corrupted LR images as the network input, we up-sample the LR images by a factor 4 using nearest interpolation.

Restoration results are evaluated with peak signal-to-noise ratio (PSNR) and structure similarity index (SSIM) (Wang et al. 2004). However, PSNR and SSIM are objective metrics that can not always be consistent with human perceptual quality. Therefore, we conduct subjective evaluation for restoration results to further investigate the effectiveness of our model. We recruited 18 subjects for subjective evaluation. We show each subject the corresponding source images prior to their evaluation for the generated images. Thus,
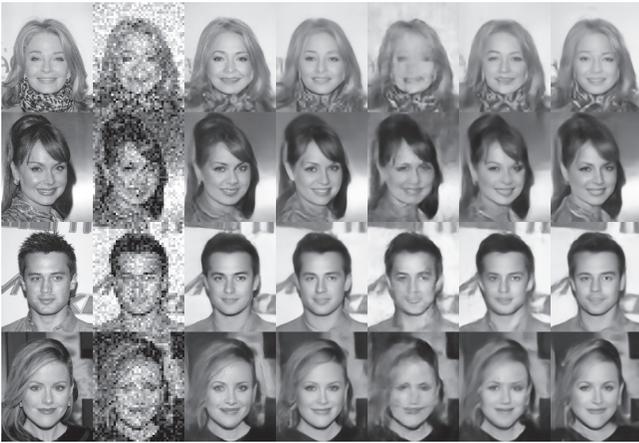
Figure 4: Restoration samples from the same scale. From left to right: ground truth, noise corrupted LR images, results from SGEN, results from SGEN-MSE, results from MEN, results from AEN, results from CEN.

Figure 5: Restoration samples from the six different scales. From left to right: ground truth, noise corrupted LR images, results from SGEN, results from SGEN-MSE, results from MEN, results from AEN, results from CEN.

they can form a general quality standard of generated images. After viewing one test image, subject gives a quality score from 1 (bad quality) to 5 (excellent quality). For each model, 288 restoration images from six different scales are evaluated, and mean opinion scores (MOS) are computed at each scale.

4.3 Comparison of different losses and ensemble methods

To investigate influence of different loss choices and verify the effectiveness of sequential ensemble structure of SGEN, we compare performance of following models:

- **SGEN**. SGEN is trained with MSE and adversarial loss.

- **SGEN-MSE**. SGEN-MSE is trained with only MSE loss.

- **MEN**. MEN (Max Ensemble Network) is just like SGEN but without any SGU during the combination of base-en/decoders, it uses max ensemble instead.

- **AEN**. AEN (Average Ensemble Network) uses average ensemble instead of SGU.

- **CEN**. AEN (Concatenate Ensemble Network) uses concatenate ensemble instead of SGU.

The objective and subjective results are shown in Table 1 and Table 2. Visual restoration samples from one scale and six scales are shown in Figure 4 and Figure 5. Only MSE loss for SGEN training provides higher PSNR and SSIM than combining MSE and adversarial loss, this result is not surprising because only minimizing MSE is equal to maximize PSNR. Smoother restoration results are also obtained by SGEN-MSE and are less visually preferred than SGEN through MOS evaluation. Comparing SGEN with other non-sequential ensemble networks, SGEN achieves better subjective scores and produces visually preferred images with more face details. Actually, because of the gate mechanism in SGU, sequential ensemble method can be viewed as a generalized ensemble method including max ensemble and
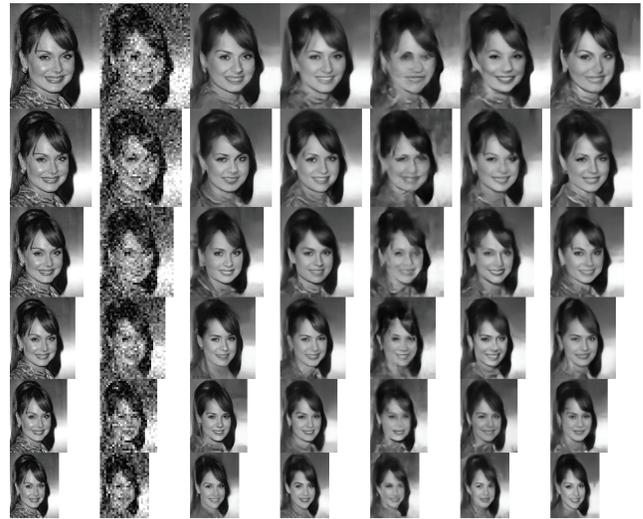
average ensemble. The CEN shows more competitive results than MEN and AEN because CEN combines base-en/decoders' information all together without any information selection. The better performance of SGEN than CEN also demonstrates the effectiveness of automatically information selection in sequential ensemble method.

4.4 Comparison with state-of-the-art algorithms

We compare the performance of SGEN with five state-of-the-art image restoration networks: SRCNN (Dong et al. 2014), SRResNet(Ledig et al. 2016), SRGAN(Ledig et al. 2016), RED-Net (Mao, Shen, and Yang 2016) and URDGN (Yu and Porikli 2016). All the networks are retrained on the same multi-scale and noisy face dataset. The quantitative results are shown in Table 3 and Table 4. The visual restoration samples from one single scale and six scales are shown in Figure 6 and Figure 7. The quantitative results confirm that our SGEN with only MSE loss achieves state-of-the-art performance in terms of PSNR and SSIM. MOS results from subjective evaluation also suggest that SGEN with adversarial training produces better perceptual quality than other algorithms. From Figure 6 and Figure 7, we can see that results from SGEN are more visually clear and contain more face details. Compared with non-ensemble structure SRCNN and URGAN, our SGEN shows better ability to handle multi-scale face restoration. Although the RED-Net, SRResNet and SRGAN try to restore image details with skip-connections that can be viewed as one kind of ensemble structure (Veit, Wilber, and Belongie 2016), the networks can not restore faces as good as SGEN due to lack of sequential ensemble method that can sequentially choose to extract high level information from corrupted input and restore more low level details.

Table 1: Average PSNR, SSIM and MOS results of networks with different loss and ensemble methods from scale 128×96 to scale 160×128. Highest scores are in bold.

	Scale 128×96			Scale 144×112			Scale 160×128		
	PSNR	SSIM	MOS	PSNR	SSIM	MOS	PSNR	SSIM	MOS
SGEN	22.37	0.6555	**4.4833**	23.08	0.6863	**4.7833**	23.61	0.7006	**4.6333**
SGEN-MSE	**23.00**	**0.6989**	4.3667	**23.60**	**0.7161**	4.5000	**24.12**	**0.7327**	4.6000
MEN	22.10	0.6485	2.0833	22.68	0.6652	2.1333	23.04	0.6807	2.0833
AEN	22.61	0.6800	3.2000	23.17	0.7015	3.4000	23.64	0.7134	3.2000
CEN	22.75	0.6945	4.0333	23.36	0.7129	4.1000	23.88	0.7261	4.2000

Table 2: Average PSNR, SSIM and MOS results of networks with different loss and ensemble methods from scale 176×144 to scale 208×176. Highest scores are in bold.

|	Scale 176×144			Scale 192×160			Scale 208×176		
	PSNR	SSIM	MOS	PSNR	SSIM	MOS	PSNR	SSIM	MOS
SGEN	24.12	0.7202	**4.6333**	24.55	0.733	4.6000	24.92	0.7429	**4.6833**
SGEN-MSE	**24.61**	**0.7501**	4.3667	**24.98**	**0.7576**	**4.6333**	**25.39**	**0.7686**	4.5833
MEN	23.50	0.6942	1.9833	23.97	0.7052	2.1167	24.46	0.7197	2.2667
AEN	24.14	0.7274	3.2333	24.62	0.7411	3.3000	24.97	0.7533	3.2667
CEN	24.42	0.7397	4.1000	24.82	0.7545	4.0500	25.20	0.7646	4.0833
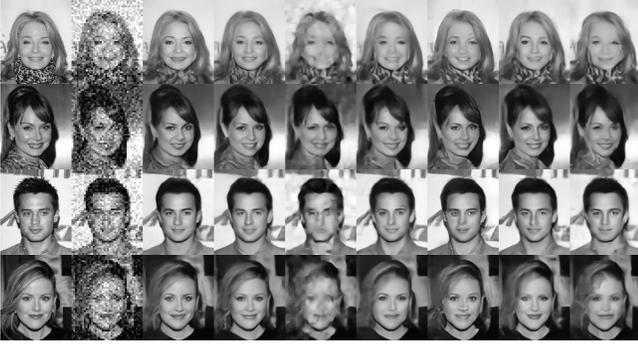
Figure 6: Restoration samples from the same scale. From left to right: ground truth, noise corrupted LR images, results from SGEN, results from SGEN-MSE, results from SRCNN, results from SRResNet, results from SRGAN, results from RED-Net, results from URDGN.

4.5 Computational Time

We train all the networks on one NVIDIA K80 GPU. The training time of the SGEN (average training time = 26 GPU-hours) is little slower than the structure with the same loss terms but without ensemble, such as URDGN (average training time = 20 GPU-hours). The sequential ensemble structure converges little slower than the single autoencoder structure is not surprising, as the SGEN learns multi-scale features and possesses a larger feature capacity. Considering the fact that our network can converge slightly more than one GPU-day, which means that our network structure is a relatively light structure and the convergence is not a problem for our network. In addition, our model only takes about $0.016s$ on average for each test image, which is well suited for real-time image processing.

4.6 Influence of N

In addition to SGEN with $N = 3$, we train other two SGENs to explore the influence of main parameter N of SGEN. The average PSNR and SSIM performance on test set, and computational cost time of SGEN with different N are shown in Table 5. Without losing the universality, only MSE loss for SGEN is used for this comparison. It is obvious that the performance of SGEN improves with the increase of N. SGEN with $N = 4$ achieves performance gain about 0.86% in terms of PSNR against SGEN with $N = 3$. However, 35% extra computational cost will also be added by SGEN with $N = 4$ compared with $N = 3$. Therefore, we choose $N = 3$ in our paper to achieve a trade-off between computational cost and performance.

5 Conclusions and Future Work

In this paper, we present a SGEN model for multi-scale face restoration from low-resolution and strong noise. We propose to aggregate multi-level base-en/decoders into the SGEN. The SGEN takes en/decoders from different levels as the sequential data. Specifically, the SGEN learns to sequentially extract high level information from low level base-encoders in bottom-up manner and sequentially restore low level details from high level base-decoders in top-down manner. Specially, we propose an elaborate SGU unit that could sequentially combine and select the multi-level information from base-en/decoders. Owing to the sequential ensemble structure, SGEN with MSE loss achieves the state-of-the-art performance of multi-scale face restoration in terms of PSNR and SSIM. Further applying adversarial loss to SGEN training, SGEN achieves the best perceptual quality according to subjective evaluation.

There are multiple aspects to explore for SGEN. First, we will apply the proposed model to other image-to-image translation tasks. Second, it is worth exploring SGEN on

Table 3: Average PSNR, SSIM and MOS comparison results with state-of-the-art algorithms from scale 128×96 to scale 160×128. Highest scores are in bold.

	Scale 128×96			Scale 144×112			Scale 160×128		
	PSNR	SSIM	MOS	PSNR	SSIM	MOS	PSNR	SSIM	MOS
SGEN	22.37	0.6555	**4.4833**	23.08	0.6863	**4.7833**	23.61	0.7006	**4.6333**
SGEN-MSE	**23.00**	**0.6989**	4.3667	**23.60**	**0.7161**	4.5000	**24.12**	**0.7327**	4.6000
SRCNN	21.72	0.5923	1.0000	22.22	0.6094	1.0667	22.69	0.6236	1.1000
SRResNet	22.73	0.6827	3.0333	23.29	0.7016	3.0667	23.81	0.7166	3.2667
SRGAN	22.29	0.6486	2.9444	22.78	0.6796	3.0000	23.43	0.6927	3.3889
RED-Net	22.77	0.6809	3.6667	23.32	0.7001	3.6333	23.83	0.7147	3.8167
URDGN	22.54	0.6688	2.8667	23.10	0.6885	3.0667	23.56	0.7044	3.0500

Table 4: Average PSNR, SSIM and MOS comparison results with state-of-the-art algorithms from scale 176×144 to scale 208×176. Highest scores are in bold.

	Scale 176×144			Scale 192×160			Scale 208×176		
	PSNR	SSIM	MOS	PSNR	SSIM	MOS	PSNR	SSIM	MOS
SGEN	24.12	0.7202	**4.6333**	24.55	0.7330	4.6000	24.92	0.7429	**4.6833**
SGEN-MSE	**24.61**	**0.7501**	4.3667	**24.98**	**0.7576**	4.6333	**25.39**	**0.7686**	4.5833
SRCNN	23.17	0.6419	1.1500	23.58	0.6548	1.2333	23.92	0.6643	1.4833
SRResNet	24.31	0.7328	3.3500	24.77	0.7458	3.0833	25.04	0.7543	3.1500
SRGAN	24.03	0.7198	3.8333	24.37	0.7297	3.9444	24.71	0.7413	3.8889
RED-Net	24.30	0.7318	3.6667	24.73	0.7461	3.6500	25.11	0.7560	3.7167
URDGN	23.98	0.7181	2.9333	24.42	0.7306	2.9500	24.67	0.7396	2.8500

Table 5: Average PSNR and SSIM performance and computational cost time of SGEN with different N.

	$N=2$	$N=3$	$N=4$
PSNR	24.13	24.28	24.49
SSIM	0.7289	0.7373	0.7463
Time	20h	26h	35h

Figure 7: Restoration samples from the six different scales. From left to right: ground truth, noise corrupted LR images, results from SGEN, results from SGEN-MSE, results from SRCNN, results from SRResNet, results from SRGAN, results from RED-Net, results from URDGN.

face video restoration in the future, which can be utilized for real-time surveillance video analysis. Third, it is also interesting to combine SGEN with face analysis techniques for end-to-end face restoration and face analysis.

6   Acknowledgement

References

Denton, E. L.; Chintala, S.; Fergus, R.; et al. 2015. Deep generative image models using a laplacian pyramid of adversarial networks. In *Advances in neural information processing systems*, 1486–1494.

Dieterich, T. G. 2000. *Ensemble Methods in Machine Learning*. Berlin, Heidelberg: Springer Berlin Heidelberg. 1–15.

Dong, C.; Loy, C. C.; He, K.; and Tang, X. 2014. Learning a deep convolutional network for image super-resolution. In *European Conference on Computer Vision*, 184–199. Springer.

Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; and Bengio, Y. 2014. Generative adversarial nets. In *Advances in neural information processing systems*, 2672–2680.

Hinton, G. E., and Salakhutdinov, R. R. 2006. Reducing the dimensionality of data with neural networks. *science* 313(5786):504–507.

Hochreiter, S., and Schmidhuber, J. 1997. Long short-term memory. *Neural computation* 9(8):1735–1780.

Jiang, J.; Hu, R.; Wang, Z.; and Han, Z. 2014. Noise robust face hallucination via locality-constrained representation. *IEEE Transactions on Multimedia* 16(5):1268–1281.

Jiang, J.; Ma, J.; Chen, C.; Jiang, X.; and Wang, Z. 2016. Noise robust face image super-resolution through smooth sparse representation. *IEEE Transactions on Cybernetics*.

Johnson, J.; Alahi, A.; and Fei-Fei, L. 2016. Perceptual losses for real-time style transfer and super-resolution. In *European Conference on Computer Vision*, 694–711. Springer.

Kingma, D., and Ba, J. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Kuncheva, L. I. 2004. *Combining pattern classifiers: methods and algorithms*. John Wiley & Sons.

Ledig, C.; Theis, L.; Huszár, F.; Caballero, J.; Cunningham, A.; Acosta, A.; Aitken, A.; Tejani, A.; Totz, J.; Wang, Z.; et al. 2016. Photo-realistic single image super-resolution using a generative adversarial network. *arXiv preprint arXiv:1609.04802*.

Lin, M.; Chen, Q.; and Yan, S. 2013. Network in network. *arXiv preprint arXiv:1312.4400*.

Liu, Z.; Luo, P.; Wang, X.; and Tang, X. 2015. Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision (ICCV)*.

Long, J.; Shelhamer, E.; and Darrell, T. 2015. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 3431–3440.

Ma, X.; Zhang, J.; and Qi, C. 2010. Hallucinating face by position-patch. *Pattern Recognition* 43(6):2224–2236.

Maas, A. L.; Hannun, A. Y.; and Ng, A. Y. 2013. Rectifier nonlinearities improve neural network acoustic models. In *Proc. ICML*, volume 30.

Mao, X.; Shen, C.; and Yang, Y.-B. 2016. Image restoration using very deep convolutional encoder-decoder networks with symmetric skip connections. In *Advances in Neural Information Processing Systems*, 2802–2810.

Nair, V., and Hinton, G. E. 2010. Rectified linear units improve restricted boltzmann machines. In *Proc. ICML*, 807–814.

Polikar, R. 2006. Ensemble based systems in decision making. *IEEE Circuits and systems magazine* 6(3):21–45.

Radford, A.; Metz, L.; and Chintala, S. 2015. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*.

Salimans, T.; Goodfellow, I.; Zaremba, W.; Cheung, V.; Radford, A.; and Chen, X. 2016. Improved techniques for training gans. In *Advances in Neural Information Processing Systems*, 2226–2234.

Sundermeyer, M.; Schlüter, R.; and Ney, H. 2012. Lstm neural networks for language modeling. In *Interspeech*, 194–197.

Sutskever, I.; Vinyals, O.; and Le, Q. V. 2014. Sequence to sequence learning with neural networks. In *Advances in neural information processing systems*, 3104–3112.

Taigman, Y.; Polyak, A.; and Wolf, L. 2016. Unsupervised cross-domain image generation. *arXiv preprint arXiv:1611.02200*.

Veit, A.; Wilber, M. J.; and Belongie, S. 2016. Residual networks behave like ensembles of relatively shallow networks. In *Advances in Neural Information Processing Systems*, 550–558.

Wang, X., and Tang, X. 2005. Hallucinating face by eigentransformation. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)* 35(3):425–434.

Wang, Z.; Bovik, A. C.; Sheikh, H. R.; and Simoncelli, E. P. 2004. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing* 13(4):600–612.

Wang, Z.; Hu, R.; Wang, S.; and Jiang, J. 2014. Face hallucination via weighted adaptive sparse regularization. *IEEE Transactions on Circuits and Systems for video Technology* 24(5):802–813.

Yoo, D.; Kim, N.; Park, S.; Paek, A. S.; and Kweon, I. S. 2016. Pixel-level domain transfer. In *European Conference on Computer Vision*, 517–532. Springer.

Yu, X., and Porikli, F. 2016. Ultra-resolving face images by discriminative generative networks. In *European Conference on Computer Vision*, 318–333. Springer.