Face Video Super-Resolution with Identity Guided Generative Adversarial Networks

Dingyi Li[1] and Zengfu Wang[1,2](✉)

[1] Department of Automation, University of Science and Technology of China,
Hefei 230027, China
lidingyi@mail.ustc.edu.cn, zfwang@ustc.edu.cn

[2] Institute of Intelligent Machines, Chinese Academy of Sciences,
Hefei 230031, China

Abstract. Faces are of particular concerns in video surveillance systems. It is challenging to reconstruct clear faces from low-resolution (LR) videos. In this paper, we propose a new method for face video super-resolution (SR) based on identity guided generative adversarial networks (GANs). We establish a two-stage convolutional neural network (CNN) for face video SR, and employ identity guided GANs to recover high-resolution (HR) facial details. Extensive experiments validate the effectiveness of our proposed method from the following aspects: fidelity, visual quality and robustness to pose, expression and illuminance variations.

Keywords: Super-resolution · Face hallucination · Identity guidance
Generative adversarial networks (GANs)

1 Introduction

Video surveillance systems grow rapidly in recent years. They appear in many streets and buildings and play an important role in safety. However, their limited resolutions and scopes prevent them from providing more important information especially on human faces. Face super-resolution (SR), also called face hallucination [17], aims at reconstructing sharp face images from low-resolution (LR) observations. The study of face SR starts from the seminal work of Baker and Kanade [1]. They introduced an example-based SR method for face image SR. Example-based methods are very suitable for face SR since face images are of relatively fixed features. Traditional face SR methods use nearest neighbor [1], principal component analysis (PCA) [18], PCA and Markov network [12], sparse representation [23]. The powerful convolutional neural network, has also been introduced into face SR [28]. Perceptual loss [6,9] and generative adversarial networks (GANs) have been employed in face SR for visually pleasant SR results [3,16,24,25]. Although plenty of algorithms have been proposed, most of these algorithms conduct experiments on aligned frontal faces under restricted

J. Yang et al. (Eds.): CCCV 2017, Part II, CCIS 772, pp. 357–369, 2017.
https://doi.org/10.1007/978-981-10-7302-1_30

illuminance. However, in real video surveillance systems, there might be pose, expression and illuminance variations. Plain learning based approaches may generate over-smoothed results when the scale factors are very large. Perceptual loss and GAN based methods produce sharp images but are difficult to guarantee high restoration fidelity.

In this paper, we introduce identity guidance to GANs. A new two-stage CNN for face video SR is proposed. Different from plain perceptual loss based methods which compute the Mean Square Error (MSE) loss between the perceptual feature maps, we employ discriminator net on the identity feature maps. Different from plain GAN based methods using raw SR and ground truth (GT) high-resolution (HR) images as the inputs of the discriminator net, we feed the identity feature maps into the discriminator net for better restoration of face elements. We compare our method with state-of-the-art SR methods and show the superiority of our method in terms of both restoration fidelity and visual quality.

We summarize our contributions as follows:

1. We propose a two-stage CNN as our generator net to facilitate visually appealing face video SR in our GAN framework.
2. We establish an identity guided GAN framework for face video SR for better fidelity and visual quality.
3. Our method is robust to pose, expression and illuminance variations in challenging videos. Our face SR results are superior compared with state-of-the-art SR methods.

The rest of the paper is organized as follows. In Sect. 2 we review related work. In Sect. 3, we illustrate the proposed method. In Sect. 4, we provide detailed experimental results and analysis. The paper is concluded in Sect. 5.

2 Related Work

Face SR has attracted much attention from researchers and companies. Baker and Kanade [1] introduced nearest neighbor in face SR. Liu *et al.* [12] employed a global linear model and a local Markov network to obtain good SR results. Ma *et al.* [13] used position patch to reconstruct faces. Jiang *et al.* [5] applied locality and sparsity constraint for least square inversion problem, leading to noise robust face SR. These methods perform well on aligned faces. When faces are not aligned, their performance drop dramatically. Yang *et al.* [22] extracted facial components in LR images and utilized component examples for learning. However, it is hard to detect facial components in extremely LR images. Deep CNNs were firstly applied in face SR by Zhou *et al.* [28]. Different from CNN based general image SR methods [6–10,14,26,27], face SR CNNs utilize the facial information. Zhu *et al.* [29] cascaded gated deep bi-network and dense face correspondence field for face SR.

GANs are also introduced in image processing tasks including face image SR to improve the visual quality of the super-resolved faces. Tuzel *et al.* [16]

proposed a deep global-local face SR net and introduced adversarial training to improve visual quality. Yu and Porikli [24] designed a discriminative generative net for face SR. Yu and Porikli [25] also presented a transformative discriminative neural network to super-resolve unaligned and very LR faces. Gulrajani *et al.* [3] compared plain GANs with Wasserstein GANs for face SR.

Although GANs have been employed in face SR, the inputs of discriminator net are simply the SR and HR images in previous methods. These methods utilize no person identity priors to distinguish one person from another, only aiming at obtaining sharper face images. In this paper, we utilize face identity features as the inputs of the discriminator net for improving the fidelity of the SR images and providing better results on facial components.

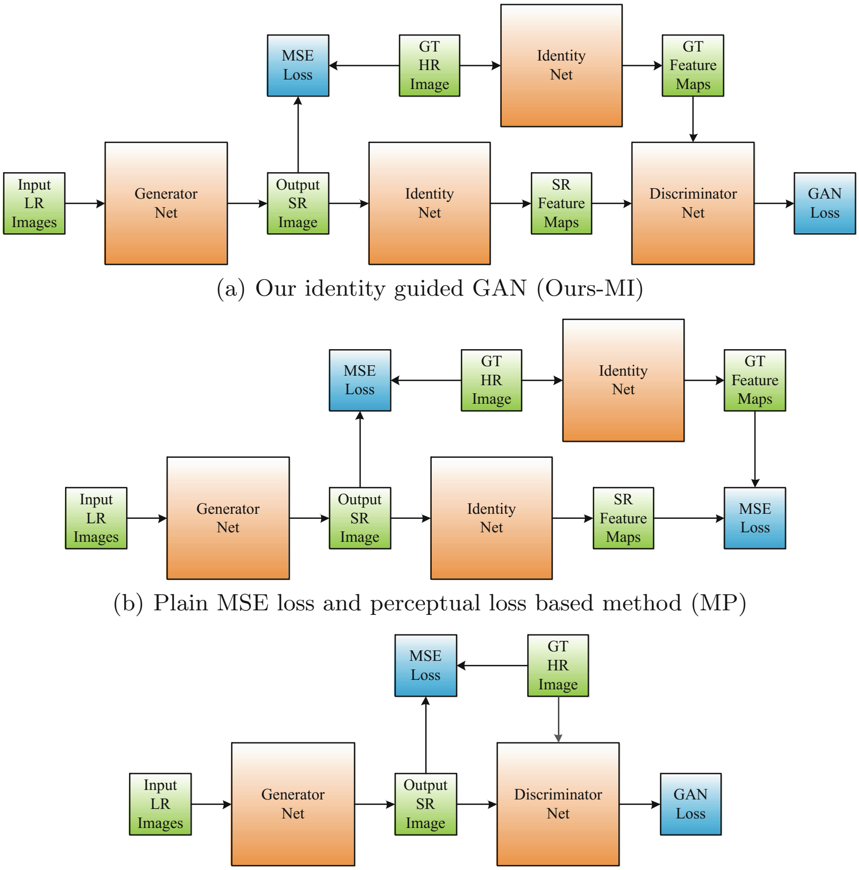
3 Our Video SR Approach

3.1 Overview of the Proposed Method

We illustrate our face video SR framework in Fig. 1(a). In our method, three networks are employed. They are the generator net, the identity net and the discriminator net. The generator net regards multiple LR images as the inputs and generates a SR image. The identity net is utilized to extract the identity feature maps of SR images and the original HR images. Our identity net applies the bottom layers of a face recognition CNN [20]. We employ the discriminator net to determine whether the SR identity feature maps look similar to the HR identity feature maps. Our method is different from plain MSE loss and perceptual loss [6,9] based SR methods as shown in Fig. 1(b) which simply employ the MSE between the SR feature maps and the HR feature maps of a pretrained VGG network [15]. Li *et al.* [11] uses identity loss to preserve identity and perceptual loss to enhance visual quality. However, these two losses are also MSE based. Apart from previous GAN based methods in Fig. 1(c) determining whether the SR images look similar to the HR images, our identity guided GAN determines whether the SR identity feature maps are similar to HR identity feature maps. Our method aims at obtaining more accurate identity features on super-resolved faces.

3.2 The Generator Net

We design a two-stage CNN as the generator net in our GAN framework for enhancing visual quality, as shown in Fig. 2. The proposed CNN is based on the state-of-the-art general video SR method named motion compensation and residual net (MCResNet) [10]. We interpolates LR inputs frames using the bicubic interpolation (Bicubic) and employs an optical flow method [2,7] for motion estimation and compensation. Then the motion compensation frames are fed into our residual video SR network. We employ two-stage deep residual learning by adding two skip-layer connections. The two stages both include a new convolutional path and an element-wise sum of the outputs of the convolutional



(a) Our identity guided GAN (Ours-MI)

(b) Plain MSE loss and perceptual loss based method (MP)

(c) Plain GAN based method (MA)

Fig. 1. Overview of our method and previous methods.

path and the interpolated centering input frame. The first stage restores high-frequency details. The blue components in Fig. 2 are the second stage of our network for denoising and illuminance adjustment. We find that although the proposed net obtains lower objective evaluation values than MCResNet using MSE loss, the new net provides better results than MCResNet in our GAN framework. Our proposed generator net adds the input centering frame twice so that the low-frequency details are better preserved in GAN training. However, the two-stage structure is a little bit worse than MCResNet with the MSE loss since the two-stage structure uses less filters.

Let $\mathbf{Y} = \{\mathbf{I}^L_{-T}, \mathbf{I}^L_{-T+1}, \cdots, \mathbf{I}^L_0, \cdots, \mathbf{I}^L_{T-1}, \mathbf{I}^L_T\}$ to be all the bicubically interpolated input frames where $\mathbf{I}^L_t$ is the t-th frame. Suppose we have N convolutional

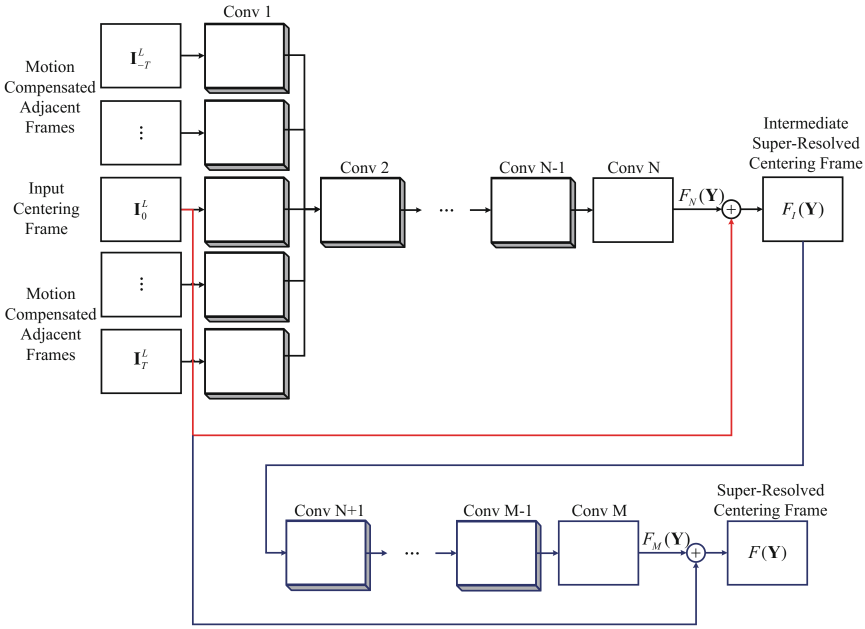


Fig. 2. The proposed generator net for face video SR. (Color figure online)

layers in the first stage and M convolutional layers in total. For the first stage we have

$$F_I(\mathbf{Y}) = F_N(\mathbf{Y}) + \mathbf{I}_0^L \tag{1}$$

where $F_I(\mathbf{Y})$ is the output of the first stage and $F_N(\mathbf{Y})$ is the output of the N-th convolutional layer. $\mathbf{I}_0^L$ is the centering frame. For the second stage we have

$$F(\mathbf{Y}) = F_M(\mathbf{Y}) + \mathbf{I}_0^L \tag{2}$$

where $F(\mathbf{Y})$ is the output of the whole network and $F_M(\mathbf{Y})$ is the output of the M-th convolutional layer. The interpolated input centering frame is added to the SR results twice for more accurate SR and easier optimization. The loss for the generator net in our GAN is

$$L_G = L_M + \lambda_A L_A \tag{3}$$

where L_M is the MSE loss and L_A is the adversarial loss. λ_A is the weight of the adversarial loss. We also have the MSE loss as

$$L_M = \frac{1}{2}||F(\mathbf{Y}, \Theta) - \mathbf{X}||_2^2 \tag{4}$$

where $F(\mathbf{Y}, \Theta)$ are the outputs of the generator net. Θ are the parameters of the generator net. $\mathbf{X}$ are the ground truth (GT) HR images. We will describe the definition of L_A in the following subsection.

3.3 The Identity Net and the Discriminator Net

For the identity net, we utilize the bottom 7 convolutional layers of one of the state-of-the-art face recognition CNN designed by Wen *et al.* [20]. The convolutional layer we use is the mid-level information extracted by the face recognition CNN. We find that high-level information at the top layers and low-level information at the bottom layers are not suitable for reconstructing facial details. Fully connected layers are not used so that the generator net is scalable to input image sizes. We use the pretrained model of [20]. The parameters of the identity net are not changed during training. The outputs of the identity net, which are the extracted identity feature maps, are directly fed into the discriminator net. For the discriminator net, we employ 4 convolutional layers with a stride of 2 and a fully connected layer. The loss of the discriminator net is

$$L_D = -log(D(I(\mathbf{X}))) - log(1 - D(I(F(\mathbf{Y})))) \tag{5}$$

where L_D is the discriminative loss. I($\mathbf{X}$) and D(I($\mathbf{X}$)) are the outputs of the identity net and the discriminator net respectively when the inputs are the original HR images. F($\mathbf{Y}$) are the outputs of generator net when the inputs are the SR images. I(F($\mathbf{Y}$)) and D(I(F($\mathbf{Y}$))) are the outputs of the identity net and the discriminator net respectively. The adversarial loss, which is back-propagated through the discriminator net and the identity net to the generator net, is defined as

$$L_A = -log(D(I(F(\mathbf{Y})))) \tag{6}$$

4 Experiments

4.1 Implementation Details

We conduct experiments on the challenging YouTube Faces Database [21]. The YouTube Faces Database contains 3,425 videos of 1,595 subjects. The videos are captured in different scenes with different illuminance. The faces are of various poses and expressions. We utilize the extracted face region videos of the first 1000 people for training, and the 3rd frame (the centering frame of the first 5 frames) of the first video of the last 95 people for testing. 181760 face images are used for training. We use Bicubic to resize all faces to a size of 80 × 80. Then we also apply Bicubic to downsample these 80 × 80 images with a scale factor of 4 to generate LR images. We conduct experiments on the Y channel in the YCbCr color space. The visual quality of the super-resolved images, which we mainly focus on, are inspected. Peak pixel-to-noise ratio (PSNR) and structural similarity (SSIM) [19] are utilized as the objective evaluations. 8 pixels on each border are eliminated when conducting the objective evaluations. We compare our method with the bicubic interpolation (Bicubic) and state-of-the-art general single image based SR methods very deep SR (VDSR) [8], denoising convolutional neural networks (DnCNN) [26] and very deep Residual Encoder-Decoder Network (RED-Net) [14], and general multi-frame based method MCResNet [10]. For VDSR, RED-Net and MCResNet, we stack 20, 30 and 20 layers respectively to achieve their best performance. All CNN based methods except for DnCNN [26] are trained on the same training dataset as ours for fair comparison using the Caffe platform [4] on a Linux workstation with an Intel i7-6700K CPU of 4.0 GHz, a Nvidia Titan X GPU and 64 GB memory. We also compare our method with face SR methods Ma's [13], structured face hallucination (SFH) [22] and Locality-constrained Representation (LcR) [5]. Since SFH is unable to detect faces on 20 × 20 low-resolution images, it provides no SR results.

In our generator net, T is 2, N is 10, M is 15. The first convolutional layer is of 64 feature maps for each frame. For layer 10 and layer 15, only 1 feature map is employed. The feature map size of the intermediate layers are 32. We also test our method with different configurations including MSE loss based (M), MSE loss and plain perceptual loss based (MP), MSE loss and plain adversarial loss based (MA), MCResNet with MSE loss and identity adversarial loss (MCRN-MI) and MSE and identity adversarial loss based (Ours-MI). For M, we employ the generator net in Fig. 2 with MSE loss. For MP, MSE loss between the SR and the HR images, and MSE loss between the SR and HR identity feature maps are computed in the structure of Fig. 1(b). For MA, we employ the framework in Fig. 1(c) as in plain GAN based SR methods. For Ous-MI we employ our identity guided GANs shown in Fig. 1(a). We use the MCResNet as the generator net for MCRN-MI. The other settings of MCRN-MI are the same as Ours-MI.

4.2 Results and Analysis

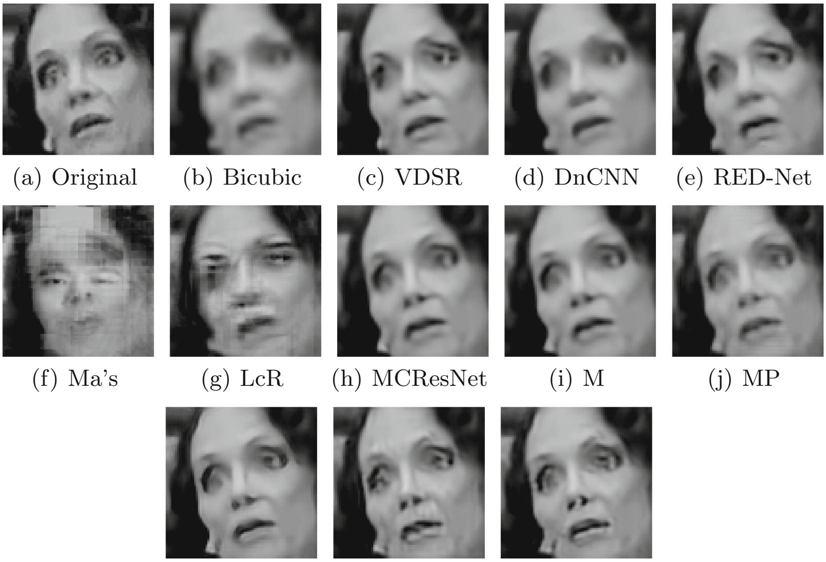
In Table 1, we show that MCResNet obtains the highest PSNR and SSIM values when trained with MSE loss only, better than our two-stage generator net which uses less filters. We also find that although our method trained with MSE loss gets slightly lower PSNR and SSIM values than MCResNet, it is more powerful to restore fine facial details when trained with both MSE loss and adversarial loss. In Table 2 we evaluate our method with different configurations objectively. MP, MA, MCRN-MI and Ours-MI get lower PSNR and SSIM values than M, but they achieve better visual quality. We find that for visual quality, Ours-MI outperforms Ma's, LcR, MP, MA and MCRN-MI. In Figs. 3, 4, 5, 6, 7, 8 and 9, we provide some of the face SR results. The results of MSE loss based methods VDSR, DnCNN, RED-Net, MCResNet, and M are over-smoothed.

Table 1. Objective evaluations of our generator net with MSE loss and other methods

Metric	Bicubic	VDSR [8]	DnCNN [26]	RED-Net [14]
PSNR	30.97	33.28	32.73	32.65
SSIM	0.8444	0.8989	0.8873	0.8868
Running time	0.001 s	0.101 s	0.004 s	0.161 s
Metric	MCResNet [10]	Ma [13]	LcR [5]	M
PSNR	34.67	23.08	27.86	34.50
SSIM	0.9241	0.5386	0.7425	0.9217
Running time	1.030 s	0.043 s	0.147 s	1.005 s

Table 2. Objective evaluations of our method with different configurations

Metric	M	MP	MA	MCRN-MI	Ours-MI
PSNR	34.50	33.14	33.18	32.90	32.81
SSIM	0.9217	0.9088	0.9083	0.8966	0.8968
Running time	1.005 s	1.007 s	1.000 s	1.017 s	0.997 s


(a) Original (b) Bicubic (c) VDSR (d) DnCNN (e) RED-Net

(f) Ma's (g) LcR (h) MCResNet (i) M (j) MP

(k) MA (l) MCRN-MI (m) Ours-MI

Fig. 3. SR results for person 1.

The results of Ma's [13] and LcR [5] are of plenty of visual artifacts since they rely on careful face alignment. The results of MP is slightly sharper than M, but looks brighter than the original image. MA can super-resolve sharp edges but is still unable to restore fine facial details especially on eyes and noses. MCRN-MI may generate fine details, but is of lower quality than Ours-MI. Meanwhile, results of MCRN-MI may contain too much noise. It is hard for MCResNet to restore fine details in our framework since it employs only one stage and loses important facial information in the deep convolutional layers. Our two-stage generator net uses two skip-layer connections to better preserve low-frequency details. We can see that our MI super-resolves fine facial components such as eyes, noses and mouths. Our identity guided GAN based method is robust to pose, expression and illuminance changes. The computational cost of each methods are shown in Tables 1 and 2. For Bicubic, Ma's and LcR, only an Intel i7-6700K CPU is used. Other methods are tested with an Intel i7-6700K CPU and a Nvidia Titan X GPU. The running time of our method is about 1 s for super-resolving one facial image, which is acceptable in real applications.

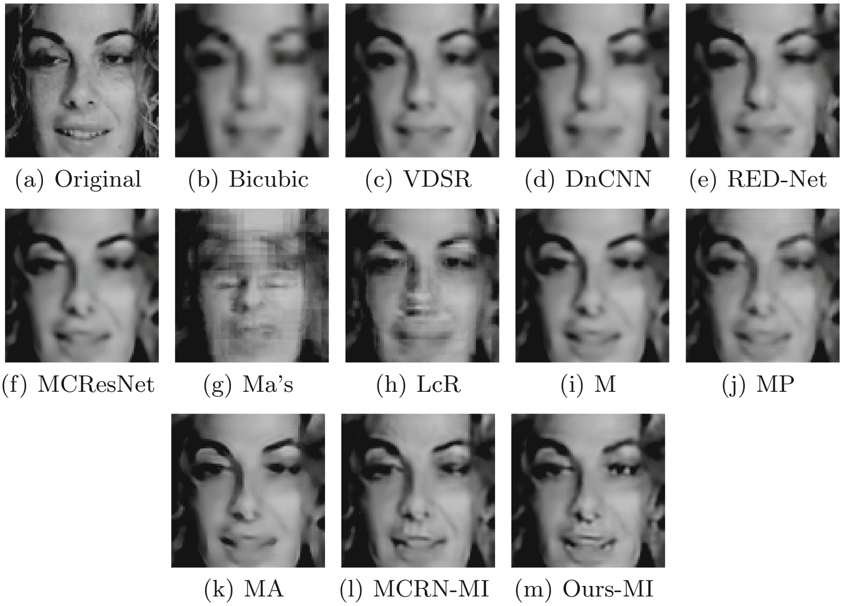


Fig. 4. SR results for person 3.

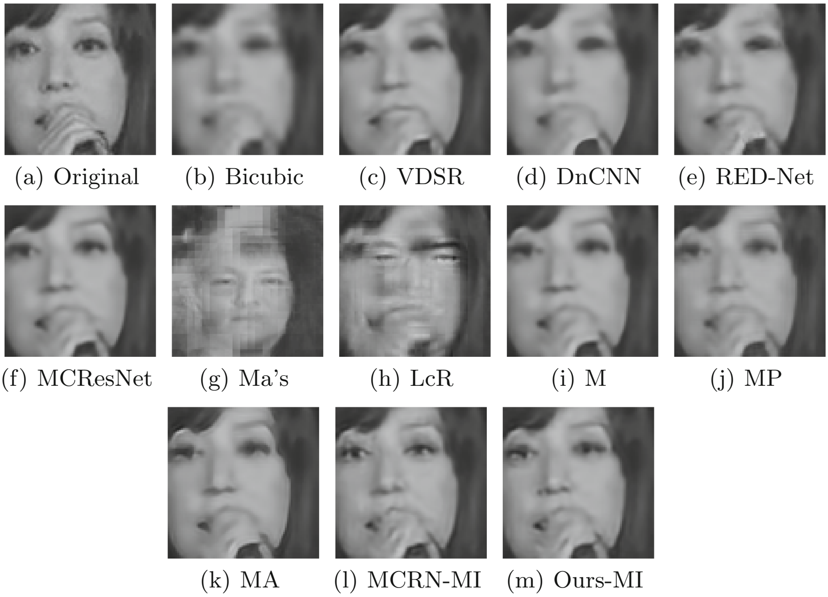


Fig. 5. SR results for person 10.

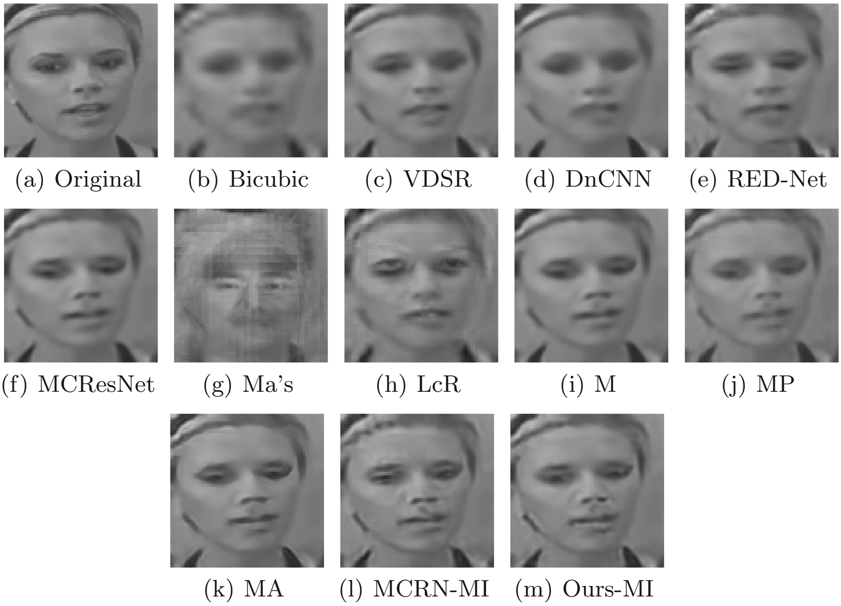


Fig. 6. SR results for person 13.

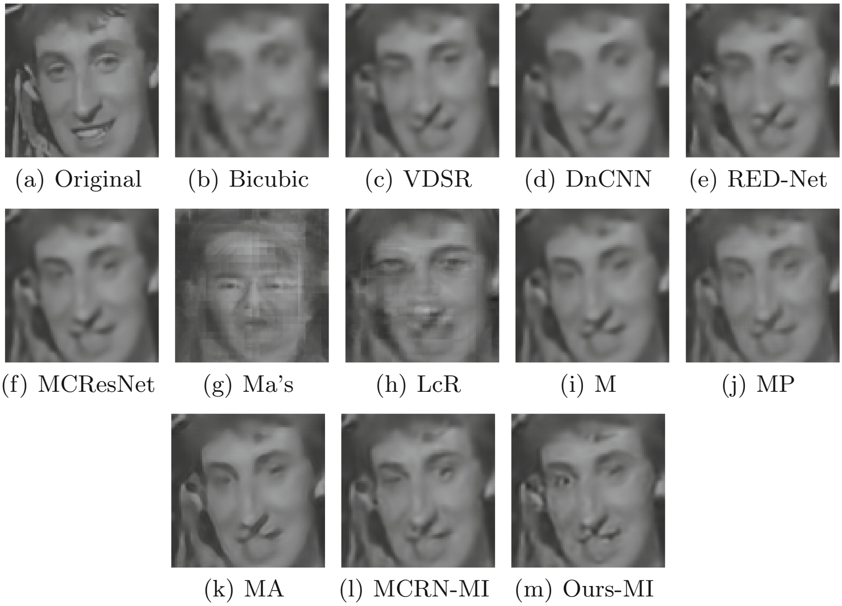
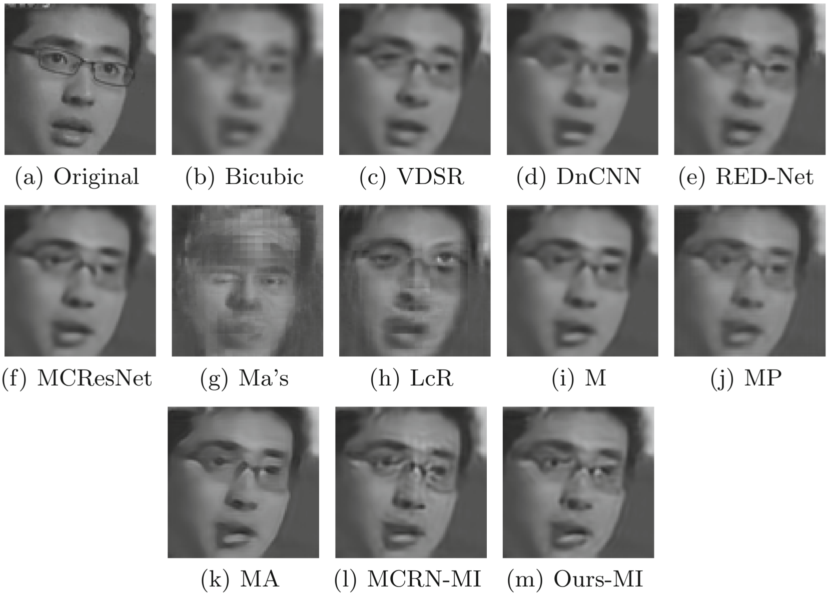


Fig. 7. SR results for person 36.

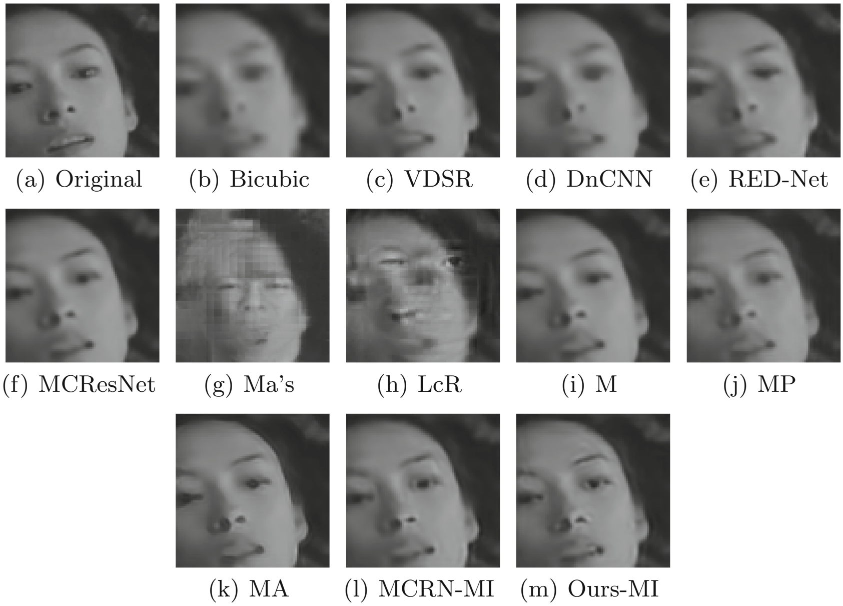


(a) Original (b) Bicubic (c) VDSR (d) DnCNN (e) RED-Net

(f) MCResNet (g) Ma's (h) LcR (i) M (j) MP

(k) MA (l) MCRN-MI (m) Ours-MI

Fig. 8. SR results for person 65.



(a) Original (b) Bicubic (c) VDSR (d) DnCNN (e) RED-Net

(f) MCResNet (g) Ma's (h) LcR (i) M (j) MP

(k) MA (l) MCRN-MI (m) Ours-MI

Fig. 9. SR results for person 90.

5 Conclusion

In this paper, we have proposed an identity guided GAN framework for face video SR. The proposed two-stage CNN generator net and the employment of the identity guidance for GANs facilitate the restoration of fine facial components. Extensive experiments have demonstrated the high fidelity and superior visual quality of our SR results. Our method is robust to pose, expression and illuminance variations. Our future work is combining face video SR with face video recognition in a joint manner.

Acknowledgements. This work was supported by the National Natural Science Foundation of China (no. 61472393).

References

1. Baker, S., Kanade, T.: Hallucinating faces. In: IEEE International Conference on Automatic Face and Gesture Recognition, pp. 83–88 (2000)
2. Drulea, M., Nedevschi, S.: Total variation regularization of local-global optical flow. In: Proceedings of the IEEE Conference on Intelligent Transportation Systems, pp. 318–323 (2011)
3. Gulrajani, I., Ahmed, F., Arjovsky, M., Dumoulin, V., Courville, A.: Improved training of wasserstein gans. arXiv preprint arXiv:1704.00028 (2017)
4. Jia, Y., Shelhamer, E., Donahue, J., Karayev, S., Long, J., Girshick, R., Guadarrama, S., Darrell, T.: Caffe: convolutional architecture for fast feature embedding. In: Proceedings of the ACM International Conference on Multimedia, pp. 675–678 (2014)
5. Jiang, J., Hu, R., Wang, Z., Han, Z.: Noise robust face hallucination via locality-constrained representation. IEEE Trans. Multimed. **16**(5), 1268–1281 (2014)
6. Johnson, J., Alahi, A., Li, F.F.: Perceptual losses for real-time style transfer and super-resolution. In: Proceedings of European Conference on Computer Vision, pp. 694–711 (2016)
7. Kappeler, A., Yoo, S., Dai, Q., Katsaggelos, A.K.: Video super-resolution with convolutional neural networks. IEEE Trans. Comput. Imaging **2**(2), 109–122 (2016)
8. Kim, J., Lee, J.K., Lee, K.M.: Accurate image super-resolution using very deep convolutional networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1646–1654 (2016)
9. Ledig, C., Theis, L., Huszar, F., Caballero, J., Aitken, A., Tejani, A., Totz, J., Wang, Z., Shi, W.: Photo-realistic single image super-resolution using a generative adversarial network. arXiv preprint arXiv:1609.04802 (2016)
10. Li, D., Wang, Z.: Video superresolution via motion compensation and deep residual learning. IEEE Trans. Comput. Imag. **3**(4), 749–762 (2017)
11. Li, M., Zuo, W., Zhang, D.: Deep identity-aware transfer of facial attributes. arXiv preprint arXiv:1610.05586 (2016)
12. Liu, C., Shum, H.Y., Freeman, W.T.: Face hallucination: theory and practice. Int. J. Comput. Vis. **75**(1), 115 (2007)
13. Ma, X., Zhang, J., Qi, C.: Hallucinating face by position-patch. Pattern Recognit. **43**(6), 2224–2236 (2010)

14. Mao, X.J., Shen, C., Yang, Y.B.: Image restoration using very deep convolutional encoder-decoder networks with symmetric skip connections. In: Proceedings of Advances in Neural Information Processing Systems, pp. 2802–2810 (2016)
15. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)
16. Tuzel, O., Taguchi, Y., Hershey, J.R.: Global-local face upsampling network. arXiv preprint arXiv:1603.07235 (2016)
17. Wang, N., Tao, D., Gao, X., Li, X., Li, J.: A comprehensive survey to face hallucination. Int. J. Comput. Vis. **106**(1), 9–30 (2014)
18. Wang, X., Tang, X.: Hallucinating face by eigentransformation. IEEE Trans. Syst. Man Cybern. C Appl. Rev. **35**(3), 425–434 (2005)
19. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: from error visibility to structural similarity. IEEE Trans. Image Process. **13**(4), 600–612 (2004)
20. Wen, Y., Zhang, K., Li, Z., Qiao, Y.: A discriminative feature learning approach for deep face recognition. In: Proceedings of European Conference on Computer Vision, pp. 499–515 (2016)
21. Wolf, L., Hassner, T., Maoz, I.: Face recognition in unconstrained videos with matched background similarity. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 529–534 (2011)
22. Yang, C.Y., Liu, S., Yang, M.H.: Structured face hallucination. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1099–1106 (2013)
23. Yang, J., Wright, J., Huang, T., Ma, Y.: Image super-resolution via sparse representation. IEEE Trans. Image Process. **19**(11), 2861–2873 (2010)
24. Yu, X., Porikli, F.: Ultra-resolving face images by discriminative generative networks. In: Proceedings of European Conference on Computer Vision, pp. 318–333 (2016)
25. Yu, X., Porikli, F.: Face hallucination with tiny unaligned images by transformative discriminative neural networks. Proceedings of AAAI Conference on Artificial Intelligence, pp. 4327–4333 (2017)
26. Zhang, K., Zuo, W., Chen, Y., Meng, D., Zhang, L.: Beyond a Gaussian denoiser: residual learning of deep CNN for image denoising. IEEE Trans. Image Process. **26**(7), 3142–3155 (2017)
27. Zhao, Y., Wang, R., Dong, W., Jia, W., Yang, J., Liu, X., Gao, W.: Gun: Gradual upsampling network for single image super-resolution. arXiv preprint arXiv:1703.04244 (2016)
28. Zhou, E., Fan, H., Cao, Z., Jiang, Y., Yin, Q.: Learning face hallucination in the wild. In: Proceedings of AAAI Conference on Artificial Intelligence, pp. 3871–3877 (2015)
29. Zhu, S., Liu, S., Loy, C.C., Tang, X.: Deep cascaded bi-network for face hallucination. In: Proceedings of European Conference on Computer Vision, pp. 614–630 (2016)