Stage-GAN with Semantic Maps for Large-scale Image Super-resolution

Zhensong Wei, Huihui Bai* and Yao Zhao
Institute of Information Science, Beijing Jiaotong University
Beijing, 100044 - China
[e-mail: zhswei@bjtu.edu.cn, hhbai@bjtu.edu.cn, yzhao@bjtu.edu.cn]
*Corresponding author: Huihui Bai

Received November 18, 2018; revised December 29, 2018; accepted February 22, 2019;
published August 31, 2019

Abstract

Recently, the models of deep super-resolution networks can successfully learn the non-linear mapping from the low-resolution inputs to high-resolution outputs. However, for large scaling factors, this approach has difficulties in learning the relation of low-resolution to high-resolution images, which lead to the poor restoration. In this paper, we propose Stage Generative Adversarial Networks (Stage-GAN) with semantic maps for image super-resolution (SR) in large scaling factors. We decompose the task of image super-resolution into a novel semantic map based reconstruction and refinement process. In the initial stage, the semantic maps based on the given low-resolution images can be generated by Stage-0 GAN. In the next stage, the generated semantic maps from Stage-0 and corresponding low-resolution images can be used to yield high-resolution images by Stage-1 GAN. In order to remove the reconstruction artifacts and blurs for high-resolution images, Stage-2 GAN based post-processing module is proposed in the last stage, which can reconstruct high-resolution images with photo-realistic details. Extensive experiments and comparisons with other SR methods demonstrate that our proposed method can restore photo-realistic images with visual improvements. For scale factor ×8, our method performs favorably against other methods in terms of gradients similarity.

Keywords: Super-resolution, Stage-GAN, Generative adversarial networks, Semantic maps, Large scaling factors

This work was supported in part by Fundamental Research Funds for the Central Universities（2018JBZ001）and Key Innovation Team of Shanxi 1331 Project (KITSX1331).

http://doi.org/10.3837/tiis.2019.08.007 **ISSN : 1976-7277**

1. Introduction

The recovery of a high resolution (HR) image from its low resolution (LR) counterpart is referred to as super-resolution (SR). SR is the topic of great interest in computer vision community and has a wide range of applications such as medical imaging [1, 2], surveillance imaging [3], satellite imaging [4] and face recognition [5].

Many SR methods have been proposed in the computer vision community. Early methods use very fast interpolation such as bicubic interpolation [6] and usually yield results with overly smooth textures. Some of the more powerful methods utilize statistical image priors [7, 8] or internal patch recurrence [9, 10]. Recently, deep learning has seen huge success in computer vision fielids such as image classification, image translation, image SR and image understanding [11,12,13,14]. For image SR, Dong et al. [15] proposed a Super-Resolution Convolutional Neural Network (SRCNN) to learn a mapping from LR to HR in an end-to-end manner. Deeper network architectures have also been shown to increase performance for SR, *e.g*. Kim et al. [16] proposed a recursive CNN that allows for long-range pixel dependencies, achieving state-of-the-art results. Pan et al. [17] proposed a general dual convolutional neural network (DualCNN) by estimating the structure and details for image SR. Residual learning has been shown to be an effective approach to achieve better performance. Lim et al. [18] used residual blocks to build a very wide network EDSR with residual scaling and a very deep MDSR. Zhang et al. [19] proposed a unified framework residual dense network (RDN) with residual dense block (RDB) for high-quality image SR.

While these SR models demonstrate promising results, there are two main issues. First, the current methods have difficulty in learning the relation between LR and HR, especially for large scaling factors. In large scaling factors, fine details of the HR image may have little or no evidence in its LR image, so the reconstructed images may not be satisfactory. Second, most of the current methods optimize the network with the mean squared error (MSE) between the reconstructed HR image and the ground truth. Since the ability of MSE loss to capture high-frequency texture details is very limited, the reconstructed HR images are often overly-smooth and have poor perceptual quality [20].

In recent years, generative modeling has been remarkable progress with the emergence of deep learning. Generative adversarial networks (GANs) [21] have emerged as a popular technique for learning generative models in computer vision. GANs consist of two networks: generator and discriminator, which are alternatively trained to compete with each other. The generator produces an image from a latent code, and the distribution of the image should ideally be indistinguishable from the distribution of the real image. GANs can provide a powerful framework for generating plausible-looking natural images with high perceptual quality. GANs enable a wide variety of application such as image generation [22, 23], image editing [24] and representation learning [25, 26].

Just as GANs learn a generative model of data, conditional GANs (cGANs) learn a conditional generative model [21]. Prior and concurrent works have conditioned GANs on text [27], discrete labels [28, 29], and images. Recent methods have achieved impressive results on image-to-image translation [12], image inpainting [30], text-to-image [31], style transfer [32] and super-resolution (SR) [33]. The key to GANs' success is the idea of adversarial training that forces the generated images to be indistinguishable from natural images. For the SR task, Ledig et al. [33] proposed a super-resolution generative adversarial network (SRGAN) for which they employ a deep residual network with skip-connection and achieved photo-realistic nature images for ×4 upscaling factor. Bulat et al. [34] proposed a High-to-Low GAN to simulate the image degradation process for creating paired low and high-resolution images, and presented a Low-to-High GAN to super-resolve the real-world low-resolution images for a specific object category, e.g. face images. Althought GAN can generate sharp and realistic images with good visual quality, the quality evaluation of the generated images is an open and difficult problem.

In this paper, we propose a novel Stage Generative Adversarial Network (Stage-GAN) conditioned on semantic maps for SR, focusing on ×4 and ×8 scaling factors. Our proposed Stage-GAN model includes three parts: semantic map generation network (Stage-0 GAN), image reconstruction network (Stage-1 GAN) and image refinement network (Stage-2 GAN). Our model uses adversarial training and the semantic information of LR image to contribute to addressing the non-linear mapping problem in large scaling factors. In the initial stage, we firstly use a pre-defined up-sampling operator (bicubic interpolation) to upscale an input LR image to the middle-resolution (MR) image. And then we use the MR image to infer the corresponding semantic map by our Stage-0 GAN. The semantic map contains some important information of the LR image for reconstructing the realistic image. In the next stage, the MR image is concatenated with the semantic map as the input of Stage-1 GAN. By conditioning on the semantic map and MR image, Stage-1 GAN learns to capture the semantic information and low-frequency information, reconstructing the HR image with more details. In last stage, we stack Stage-2 GAN to refine the Stage-1 results, yielding the photo-realistic HR images. We validate the proposed approach and compare our performance against previous works including [15, 33, 35]. Extensive experiments show that our proposed model generates high-resolution images with better perceptual quality.

The remainder of this paper is organized as follows. In Section 2, Stage-GAN is presented in detail. The experimental results and comparisons with other methods are demonstrated in Section 3. The conclusion of this paper is presented in Section 4.

2. Proposed Networks

In this section, we present each component of our model in detail. As shown in **Fig. 1**, our proposed Stage-GAN consists of three parts: semantic map generation network (Stage-0 GAN), image reconstruction network (Stage-1 GAN) and image refinement network (Stage-2 GAN). The Stage-0 GAN generates a semantic map from the input image I_{S_0}. Then the image I_{S_0} is concatenated with the corresponding semantic map I_{sem}, which serves as the input for Stage-1 GAN. Conditioned on Stage-1 results, the Stage-2 GAN refines the details of the results and sharpens the edges of objects, yielding a more realistic high-resolution image.

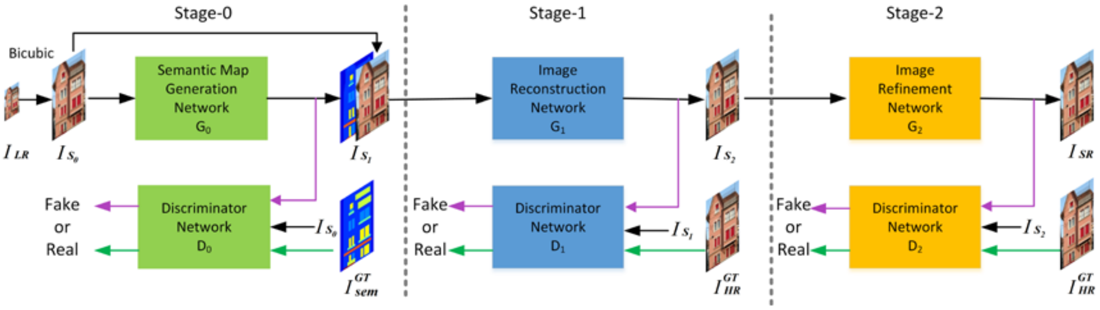

Stage-0
Stage-1
Stage-2
Bicubic
Semantic Map Generation Network G_0
Image Reconstruction Network G_1
Image Refinement Network G_2
I_{LR}
I_{S_0}
I_{S_1}
I_{S_2}
I_{SR}
Fake or Real
Discriminator Network D_0
Discriminator Network D_1
Discriminator Network D_2
I_{S_0}
I_{S_1}
I_{S_2}
I_{sem}^{GT}
I_{HR}^{GT}
I_{HR}^{GT}

Fig. 1. The architecture of our proposed Stage-GAN.

2.1 Semantic Map Generation Network (Stage-0 GAN)

For the task of image SR, to reconstruct HR image while preserving photo-realistic image details, we combine the benefits of semantic map and adversarial training to super-resolve the ill-posed problem. Different from most methods directly predicting HR image from LR image, we firstly generate a corresponding semantic map with our Stage-0 network. Given an input LR image I_{LR} , bicubic interpolation algorithm can be employed to obtain the middle-resolution (MR) image I_{S_0} . Then the MR image I_{S_0} is fed into the Stage-0 GAN to generate the corresponding semantic map I_{sem},

$$I_{sem} = F_{G_0}(I_{S_0}) = F_{G_0}(F_{bic}(I_{LR})) \tag{1}$$

where $F_{G_0}(\cdot)$ denotes the semantic map generation function and $F_{bic}(\cdot)$ denotes the bicubic interpolation operator. Specifically, we select U-Net [36] as our generator model due to its simplicity and effectiveness in semantic segmentation tasks, as shown in **Fig. 2**. Basically, U-Net is a Fully Convolutional Network [37]. It contains a series of down-sampling layers followed by a series of up-sampling layers. The feature maps are cropped and copied from down-sampling layers to up-sampling layers. It is noted that we remove the cropping and copying unit from the basic U-Net model and use only concatenation operations, yielding an improved architecture that results in better performance. And we modify the padding scheme to make the input and output of the network have the same spatial size. As shown in **Fig. 2**, the network consists of two main parts: the convolutional encoding and decoding units. The basic convolution operations are performed followed by Batch normalization [38] and ReLU activation in both parts of the network, except that the last one uses *tanh* activation. In the encoding unit, the convolution layers with kernel size 4×4, stride 2 are designed to capture useful information. The ReLU in the encoder is leaky. In the decoding phase, the de-convolution operation with kernel size 4×4, stride 2 is performed to up-sample the feature maps. We use the skip connections to concatenate feature maps from the encoding unit to the decoding unit. Some examples of generated semantic maps are shown in **Fig. 5**.

For the discriminator D_0 , as shown in **Fig. 4**, the MR image I_{S_0} is firstly concatenated along the channel dimension with the output semantic map I_{sem} as the input of 'fake' discriminator. Meanwhile, the MR image I_{S_0} is concatenated along the channel dimension with the ground truth semantic map I_{sem}^{GT} as the input of 'real' discriminator. The concatenated results are fed through three convolution layers of down-sampling with kernel size 4×4, stride 2. The last two layers with kernel size 4×4, stride 1 are used. Here, the output has 30×30 spatial dimension. Finally, this discriminator tries to determine if each 30×30 patch in an output image is 'real' or 'fake'. We run this discriminator convolutionally across the concatenated map, averaging all responses to provide the ultimate output of the discriminator D_0.

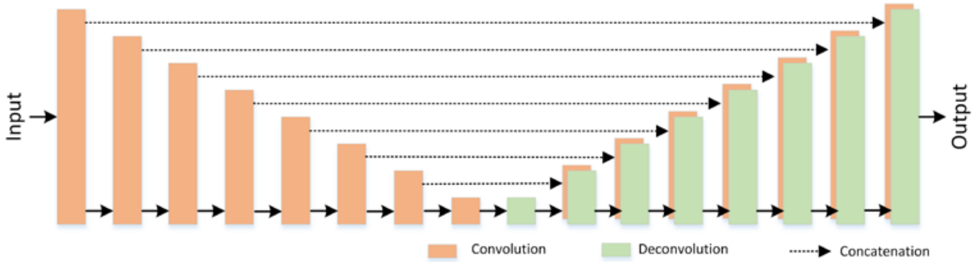


Fig. 2. The architecture of improved U-net generator.

2.2 Image Reconstruction Network (Stage-1 GAN)

We now present our Stage-1 GAN, which is used to learn a mapping from the MR image and corresponding semantic map to the desired HR image. Our Stage-1 GAN is conditioned on Stage-0 results and corresponding MR images to generate HR images. In the Stage-1 GAN, the Stage-0 result I_{sem} and the corresponding MR image I_{S_0} are delivered to the generator G_1 for image reconstruction,

$$I_{S_2} = F_{G_1}(I_{S_1}) = F_{G_1}([I_{S_0}, I_{sem}]) \tag{2}$$

where the input I_{S_1} to Stage-1 GAN is a concatenation of the MR image I_{S_0} and corresponding semantic map I_{sem}. I_{S_2} is the reconstructed image from the input $[I_{S_0}, I_{sem}]$ by the generator G_1. An $F_{G_1}(\cdot)$ denotes the image reconstruction function. The semantic map I_{sem} contains the semantic information of the LR image and the MR image I_{S_0} remains the low-frequency information of the LR image.

In order to combine the semantic information and low-frequency information, we use the U-Net as our reconstruction network, as shown in **Fig. 2**. Although GAN-based synthesized images from semantic maps are visually appealing, their details can be quite different from the original images. To obtain high-quality reconstructed images, we use the MR images and corresponding semantic maps as the input of Stage-1 GAN. In our proposed Stage-GAN, as shown in **Fig. 1**, to formulate the SR problem by considering both the semantic information and low-frequency information, the results $[I_{S_0}, I_{sem}]$ are fed into the generator G_1 of Stage-1 GAN.

For the discriminator D_1, as shown in **Fig. 4**, the conditional input I_{S_1} is firstly concatenated along the channel dimension with the reconstruction image I_{S_2} as the input of 'fake' discriminator. Meanwhile, the conditional input I_{S_1} is concatenated along the channel dimension with the ground truth HR image I_{HR}^{GT} as the input of 'real' discriminator. The concatenated results are fed through a series of convolution layers until it has 30×30 spatial dimension. The discriminator D_1 also try to determine if each 30×30 patch is 'real' or 'fake'. Details of the structure are discussed in the Stage-0 GAN parts.

2.3 Image Refinement Network (Stage-2 GAN)

The results of the Stage-1 GAN may not be satisfactory in visual quality. Some details in the reconstructed images are omitted in the Stage-1 GAN, which is vital for generating photo-realistic images. In order to improve the quality of reconstructed images, an effective post- processing module is designed, as shown in **Fig. 3**. We stack an image refinement network at the output of the Stage-1 GAN as a refiner, named Stage-2 GAN. Our Stage-2 GAN is conditioned on Stage-1 GAN results to generate high-resolution images with more photo-realistic details. In the Stage-2 GAN, the Stage-1 result I_{S_2} is fed into the generator G_2 for image refinement,

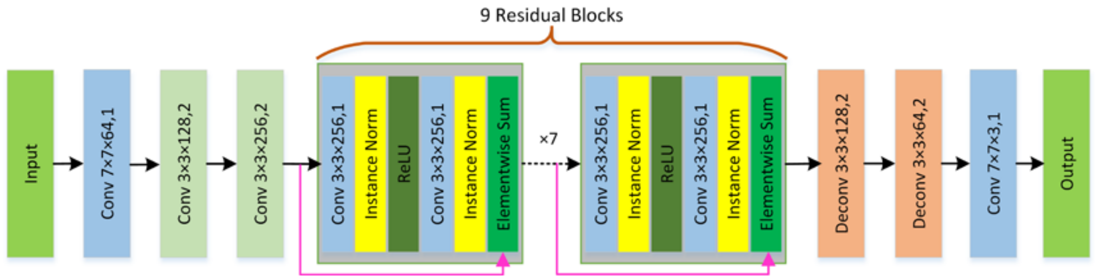


Fig. 3. Illustration of image refinement network. Convolution parameters are denoted as kernel height × kernel width × number of feature maps, stride for each convolutional layers.

$$I_{SR} = F_{G_2}(I_{S_2}) \tag{3}$$

where the input I_{S_2} to Stage-2 GAN is the result of Stage-1 GAN. And $F_{G_2}(\cdot)$ denotes the image refinement function. We design our Stage-2 generator as an encoder-decoder network with residual blocks [39]. The Stage-2 GAN completes the details of reconstructed images to generate photo-realistic images. In the Stage-2 GAN in **Fig. 3**, the first layer with kernel size 7×7, stride 1 is designed to capture more image information. The next two convolution layers (encoder) with kernel size 3×3, stride 2 are performed to down-sample the feature maps. Then, the encoder image features are fed into nine residual blocks, which are designed to learn the differences between the input image I_{S_2} and the ground truth HR image I_{HR}^{GT}. To keep the output size of decoder same with the input, the de-convolution layers (decoder) with kernel size 3×3, stride 2 are performed to up-sample the feature maps. We use the kernel size 7×7, stride 1 in the last layer. Such a generator with residual blocks is able to refine the details and sharp the edges of the objects, generating photo-realistic HR image.

For the discriminator D_2, as shown in **Fig. 4**, the reconstructed image I_{S_2} is concatenated along the channel dimension with the reconstruction image I_{SR} as the input of 'fake' discriminator. Meanwhile, the reconstructed image I_{S_2} is concatenated along the channel dimension with the ground truth HR image I_{HR}^{GT} as the input of 'real' discriminator. The concatenated results are fed through a series of convolution layers until it has 30×30 spatial dimension. The discriminator D_2 also try to determine if each 30×30 patch is 'real' or 'fake'. Details of the structure are discussed in the Stage-0 GAN parts.

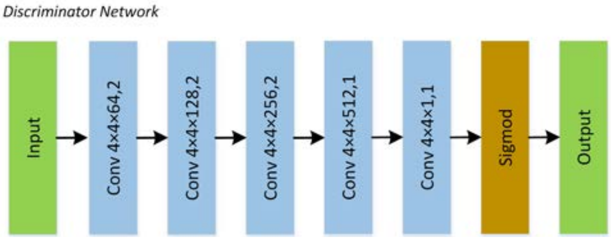


Fig. 4. The details of our discriminator architecture with corresponding kernel size, number of feature maps and stride indicated each convolutional layers.

2.4 Loss Function

Our proposed Stage-GAN consists of three parts: semantic map generation network (Stage-0 GAN), image reconstruction network (Stage-1 GAN) and image refinement network (Stage-2 GAN). Each part is based on the conditional GANs. The conditional GANs learn an adversarial loss that tries to determine if the output image is 'real' or 'fake', while simultaneously train a generative model to minimize this object.

In Stage-0 GAN, the adversarial loss function can be expressed as,

$$L_{cGAN}(G_0,D_0)=\min_{G_0}\max_{D_0} E[\log D_0(I_{sem}^{GT},I_{S_0})]+E[\log(1-D_0(G_0(I_{S_0}),I_{S_0}))] \tag{4}$$

where I_{S_0} is the MR image of LR image and I_{sem}^{GT} is the ground truth semantic map. Conditioned on the MR image I_{S_0}, Stage-0 GAN trains the discriminator D_0 and the generator G_0 by alternatively maximizing D_0 and minimizing G_0. We provide noise in form of dropout rather than gaussian noise.

In Stage-1 GAN, the adversarial loss function can be expressed as,

$$L_{cGAN}(G_1,D_1)=\min_{G_1}\max_{D_1} E[\log D_1(I_{HR}^{GT},I_{S_1})]+E[\log(1-D_1(G_1([I_{S_0},I_{sem}]),I_{S_1}))] \tag{5}$$

where I_{S_1} is a concatenation of the MR image I_{S_0} and corresponding semantic map I_{sem} and I_{HR}^{GT} is the ground truth HR image. Conditioned on the MR image I_{S_0} and corresponding semantic map I_{sem}, Stage-1 GAN trains the discriminator D_1 and the generator G_1 by alternatively maximizing D_1 and minimizing G_1. Here, we also provide noise in form of dropout rather than gaussian noise.

In Stage-2 GAN, the adversarial loss function can be expressed as,

$$L_{cGAN}(G_2,D_2)=\min_{G_2}\max_{D_2} E[\log D_2(I_{HR}^{GT},I_{S_2})]+E[\log(1-D_2(G_2(I_{S_2}),I_{S_2}))] \tag{6}$$

where I_{S_2} is the output of Stage-1 GAN and I_{HR}^{GT} is the ground truth HR image. Conditioned on the reconstructed image I_{S_2}, Stage-2 GAN trains the discriminator D_2 and the generator G_2

by alternatively maximizing D_2 and minimizing G_2 .

For image SR, most methods of supervised SR algorithms optimize the network with the L_2 distance between the reconstructed image and the ground truth image. Since the L_2 loss fails to capture perceptually relevant differences, such as high texture detail, the reconstruct-ed images are often overly smooth. Especially in large scaling factors, the reconstructed images look blurry, which is not close to human visual perception. In order to address the problem, we use L_1 distance rather than L_2 distance.

In our Stage-GAN, the L_1 loss function of Stage-0 GAN, Stage-1 GAN and Stage-2 GAN can be respectively formulated as,

$$L_1(G_0) = E[\| I_{sem}^{GT} - I_{sem} \|_1] \tag{7}$$

$$L_1(G_1) = E[\| I_{HR}^{GT} - I_{S_2} \|_1] \tag{8}$$

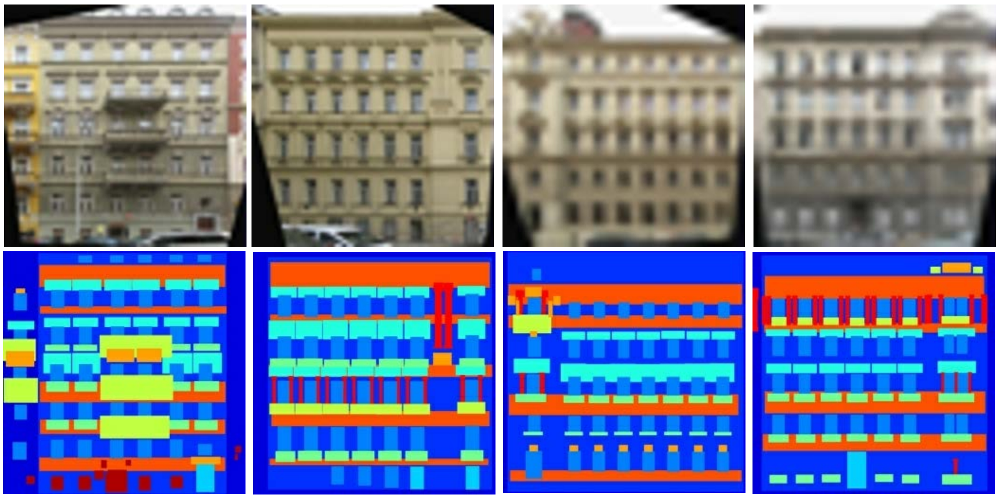
$$L_1(G_2) = E[\| I_{HR}^{GT} - I_{SR} \|_1] \tag{9}$$

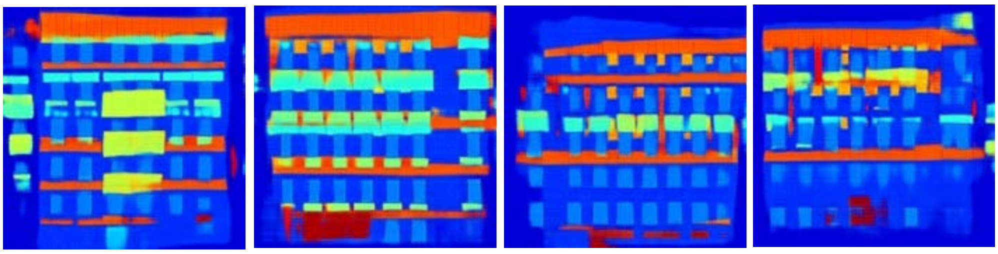
where $L_1(G_i)$ represents the L_1 loss function of Stage- i GAN, i=0,1,2 . I_{sem} is the semantic map generated by Stage-0 GAN. I_{S_2} is the reconstructed image by Stage-1 GAN. I_{SR} is the refined image by Stage-2 GAN. I_{sem}^{GT} is the ground truth semantic map and I_{HR}^{GT} is the ground truth HR image.

Finally, the loss function for our Stage-GAN can be represented as,

$$\begin{aligned} L_{loss} &= L_{cGAN}(G,D) + \lambda \cdot L_1(G) \\ &= L_{cGAN}(G_0,D_0) + L_{cGAN}(G_1,D_1) + L_{cGAN}(G_2,D_2) \\ &\quad + \lambda \cdot [L_1(G_0) + L_1(G_1) + L_1(G_2)] \end{aligned} \tag{10}$$

where L_{loss} represents the full loss of our Stage-GAN and λ is a parameter that balances the adversiral loss and L_1 loss. In this work, we set the parameter λ=100 for our experiment, which was used in [12]. The generator is tasked to not only fool the discriminator but also to generate near the ground truth output image in an L_1 sense.





(a) (b) (c) (d)

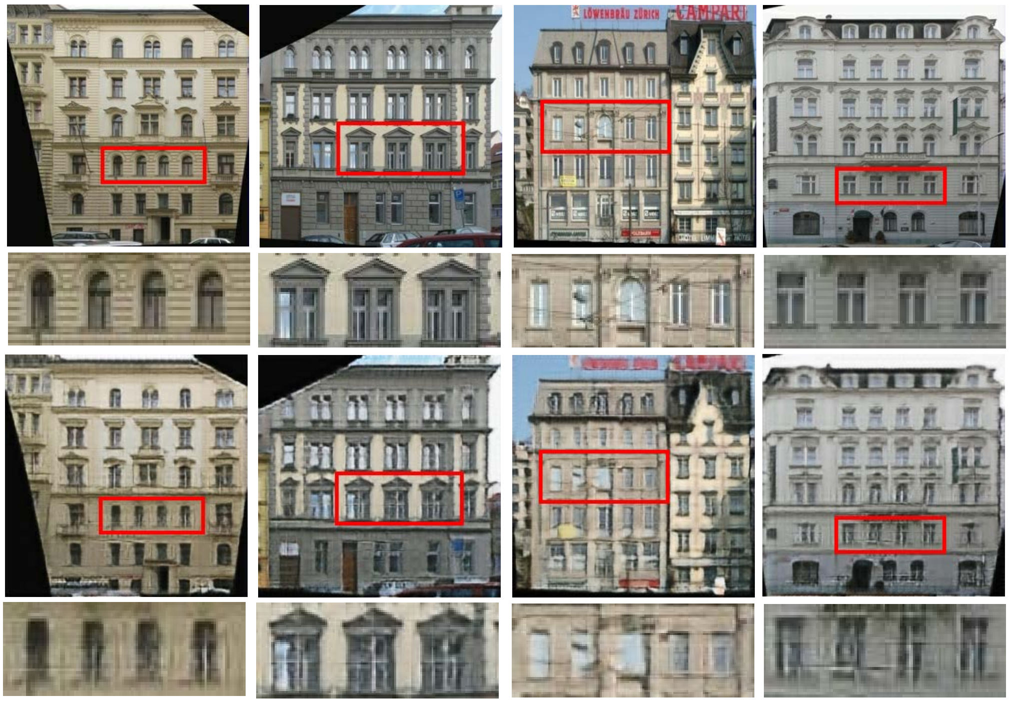
Fig. 5. Example images with semantic maps. Top: the MR images; middle: the ground truth semantic maps; bottom: the semantic maps generated by Stage-0 GAN. (a) and (b) are for ×4 scale factor, (c) and (d) are for ×8 scale factor.

3. Experimental Results and Analysis

In this section, we evaluate the performance of our model. Here, we first describe implementation details of this work and evaluation metrics. Then we analysis important components of our proposed Stage-GAN. Finally, the proposed method is compared with several SR methods.

3.1 Implementation Details and Evaluation Metrics

Our model is trained in a supervised fashion on pairs of images and semantic maps. Such pairs are provided with semantic segmentation datasets. In this work, we use the CMP Facades dataset [40], which consists of just 400 images for training. We use the CMP Facades validation set for testing, which consists of 100 images. The facades are from different cities around the world and diverse architectural styles. We sample the original images to 256×256



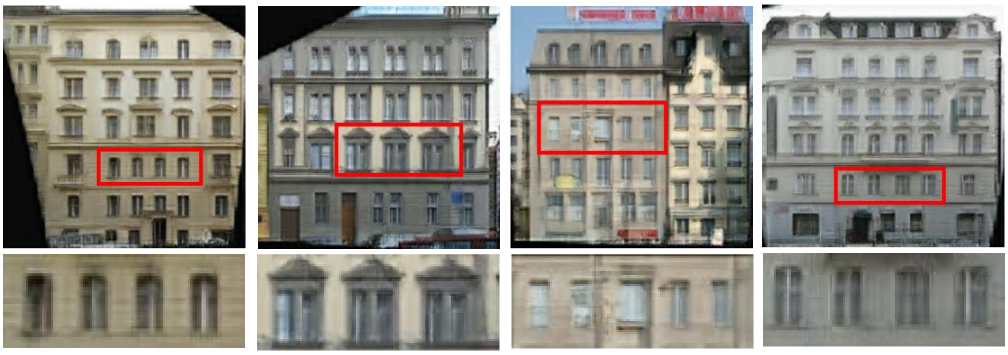
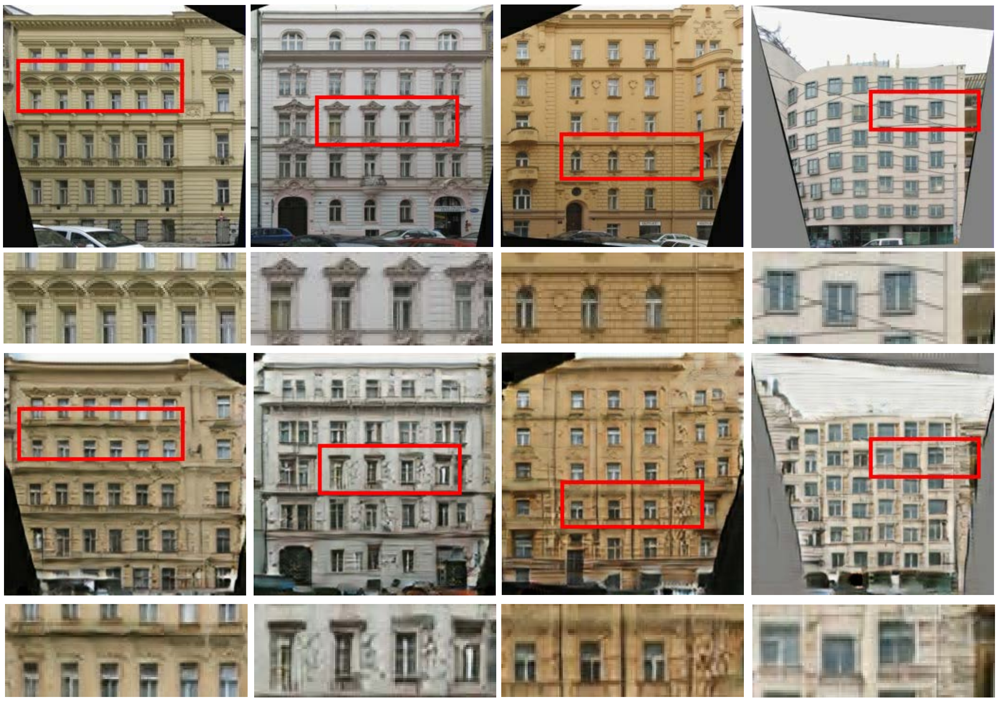


Fig. 6. Visual comparisons on several images with the scaling factor ×4. Top: the ground truth HR images; middle: the images generated by Stage-1 GAN without semantic maps; bottom: the images generated by Stage-1 GAN with semantic maps.

resolution and scale the range of the images to [0,1] for our experiments. We obtain the LR images by down-sampling the original HR images using bicubic interpolation [6] with scaling factors of ×4 and ×8. In the U-Net, all ReLUs in the encoder are leaky, with slope 0.2, while in the decoder are not leaky. Our refinement network uses a framework with 9 residual blocks, as illustrated in **Fig. 3**. For the discriminator network, we use 30×30 Patch-GAN, as illustrated in **Fig. 4**. The convolution operations are performed followed by Batch normalization [38] and leaky ReLU activation with slope 0.2, except the last one. In our training process, we use Adam solver [41] with a mini-batch size of 1 and a momentum parameter of 0.5. The weights are initialized from a Gaussian distribution with mean 0 and standard deviation 0.02. Learning rate is initially set to 0.0002 and then linearly decay to zero every 100 epochs. All reported PSNR(dB) [42] and SSIM [43] measures are calculated on Y-channel of images.



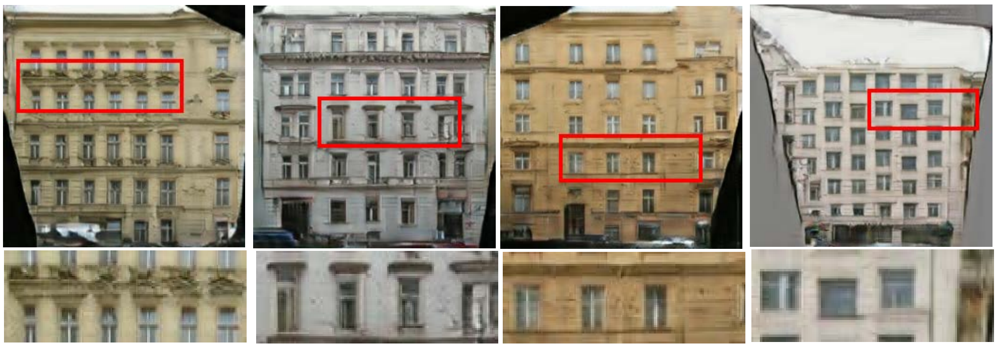
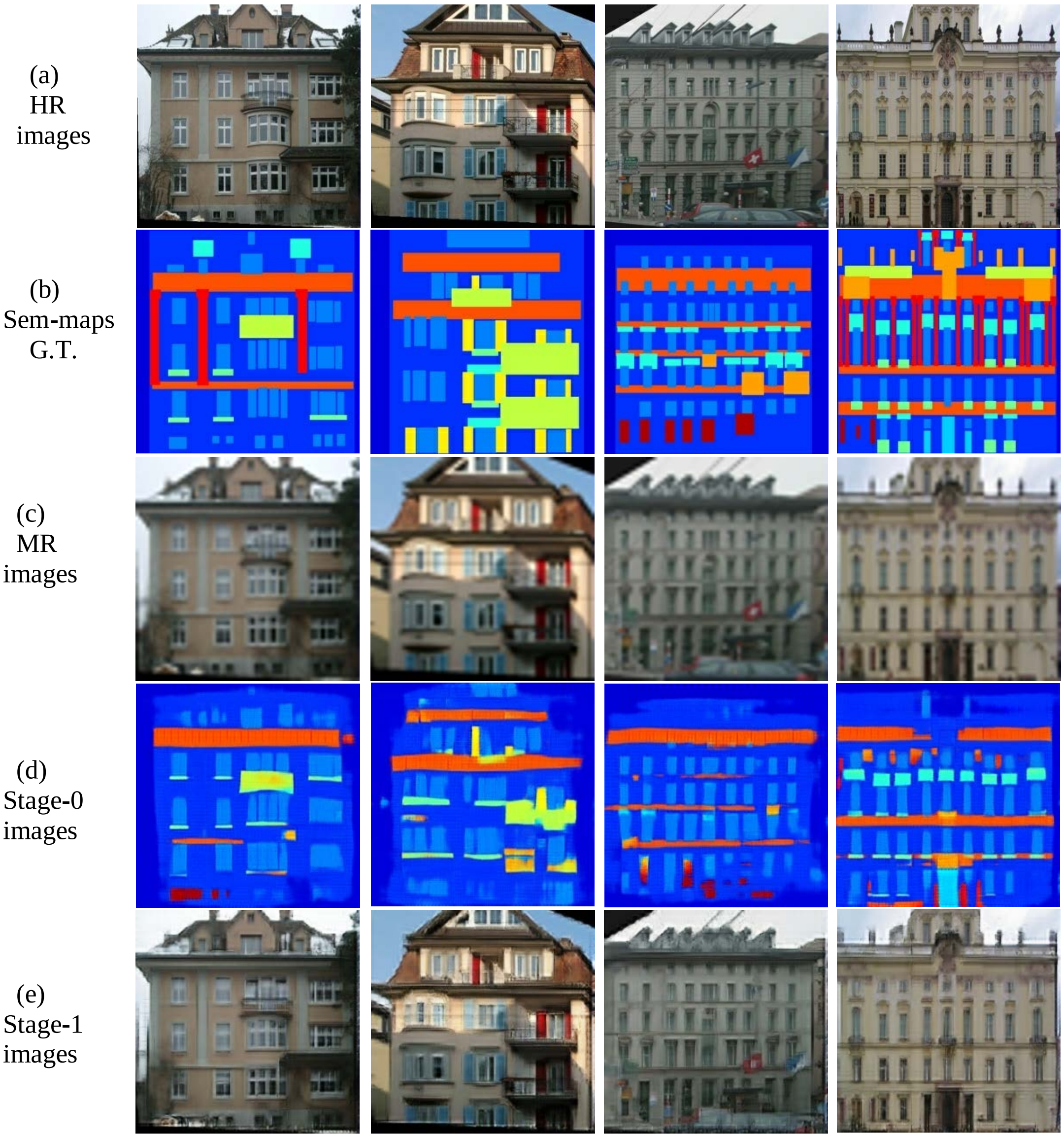


Fig. 7. Visual comparisons on several images with the scaling factor ×8. Top: the ground truth HR images; middle: the images generated by Stage-2 GAN conditioned on the reconstructed images without semantic map; bottom: the images generated by Stage-2 GAN conditioned on the reconstructed images with semantic map.



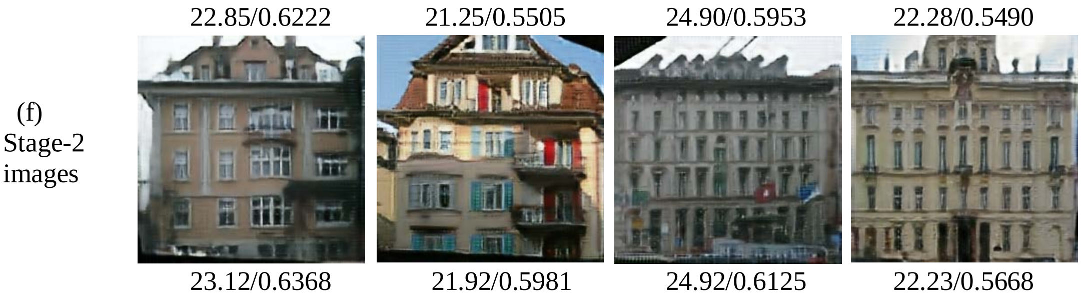
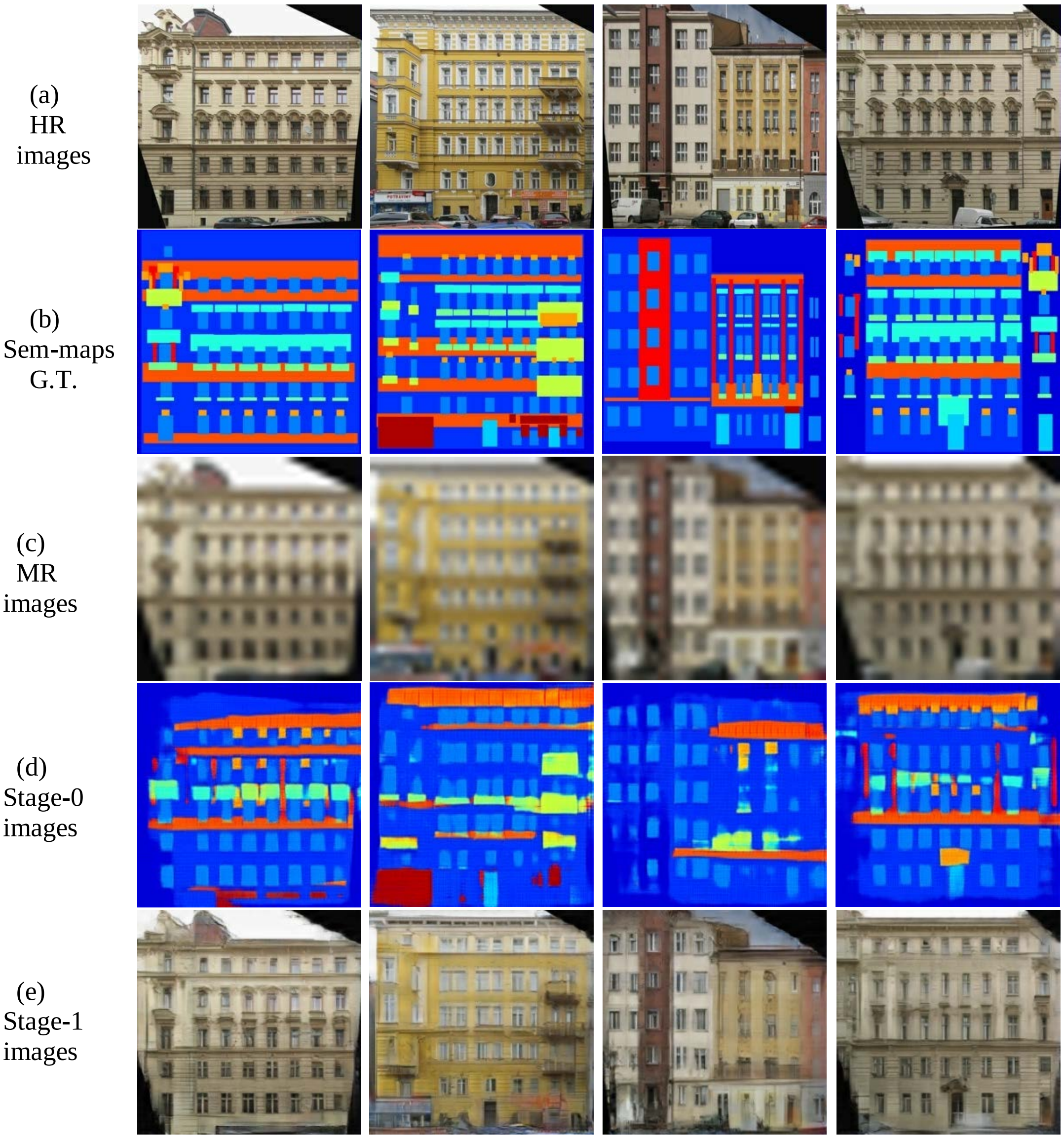


Fig. 8. Visual comparisons on several images generated by our Stage-GAN with the scaling factor ×4. Each row lists the HR images, the ground truth semantic maps, the MR images, semantic maps generated by Stage-0 GAN, images reconstructed by Stage-1 GAN and images refined by Stage-2 GAN. The images generated by Stage-2 GAN have much cleaner and sharper details than the output images of Stage-1 GAN. Corresponding PSNR(dB) [42] and SSIM [43] are shown in bottom.



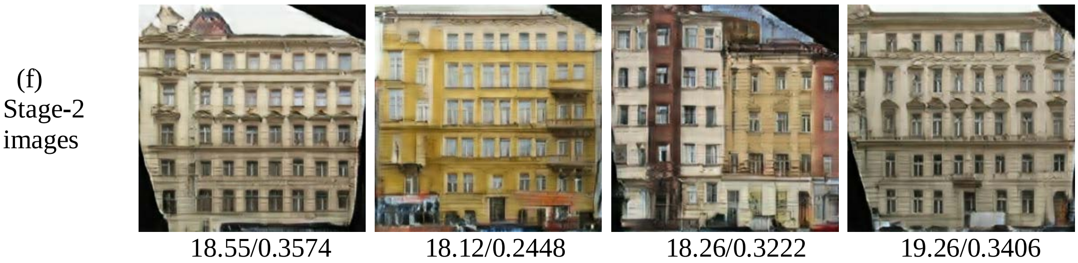


Fig. 9. Visual comparisons on several images generated by our Stage-GAN with the scaling factor ×8. Each row lists the HR images, the ground truth semantic maps, the MR images, semantic maps generated by Stage-0 GAN, images reconstructed by Stage-1 GAN and images refined by Stage-2 GAN. The images generated by Stage-2 GAN have much sharper details than the images generated by Stage-1 GAN. Corresponding PSNR(dB) [42] and SSIM [43] are shown in bottom.

For generative models (*e.g.*, GAN), the quality evaluation of the generated images is an open and difficult problem. The traditional metrics used to evaluate the SR images are PSNR [42] and SSIM [43], both of which have been found to correlate poorly with human assessment of visual quality. We therefore emphasize that the goal of our experiments is not to achieve state-of-the-art PSNR or SSIM results, but instead to generate HR images with high perceptual quality. For an image, the gradients can convey important visual information, which are crucial to scene understanding. So, we use a new image quality assessment scheme, with emphasis on gradient similarity (GSM) [44], for measuring the change in contrast and structure in images.

3.2 Component Analysis

In this work, our proposed Stage-GAN includes three parts: semantic map generation network (Stage-0 GAN), image reconstruction network (Stage-1 GAN) and image refinement network (Stage-2 GAN). As illustrated in **Fig. 1**, in Stage-1 GAN, the generators G_0 can generate semantic maps from the input MR images, which preserve the semantic information of LR images. In the training phase, we first generate the semantic maps and the examples are in **Fig. 5**.

For the large scaling factors, fine details of the HR images may have little or no evidence in its LR images. It is difficult to learn the non-linear mapping from LR to HR images. In Stage-1 GAN, considering that semantic information can contribute to producing correspond-

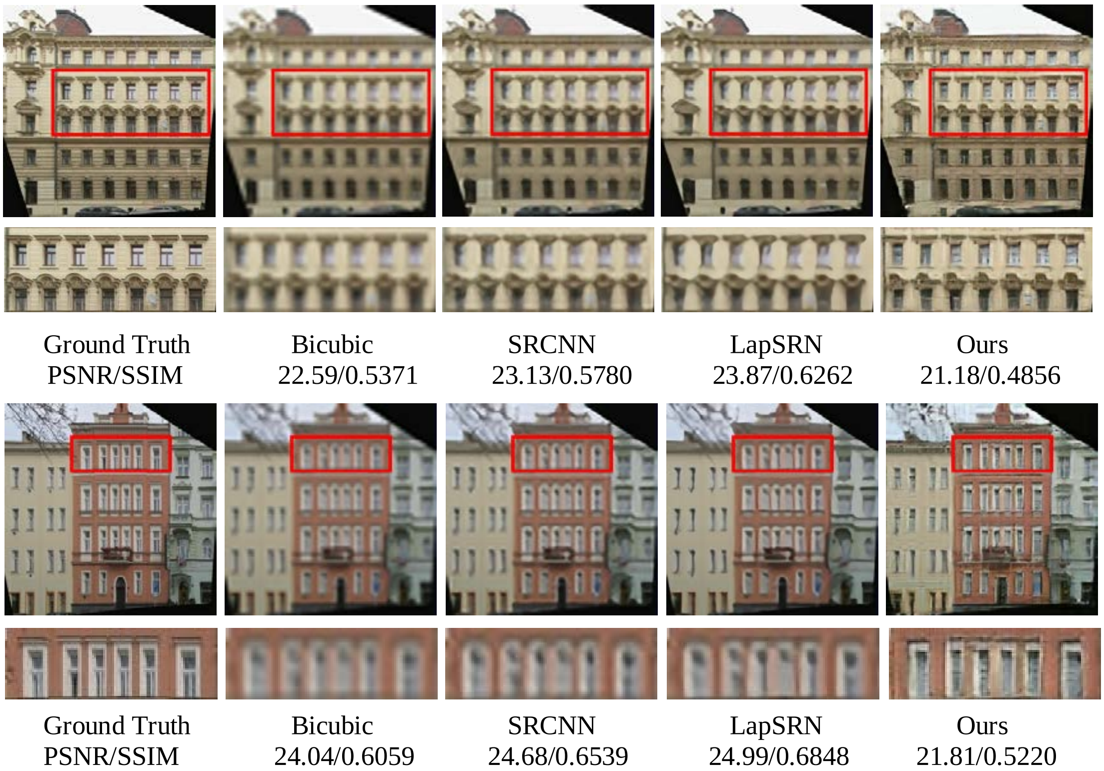
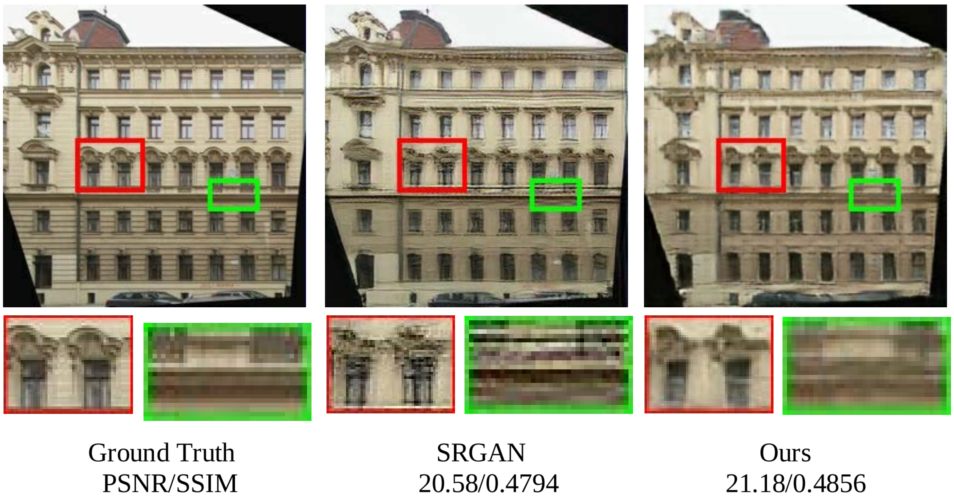
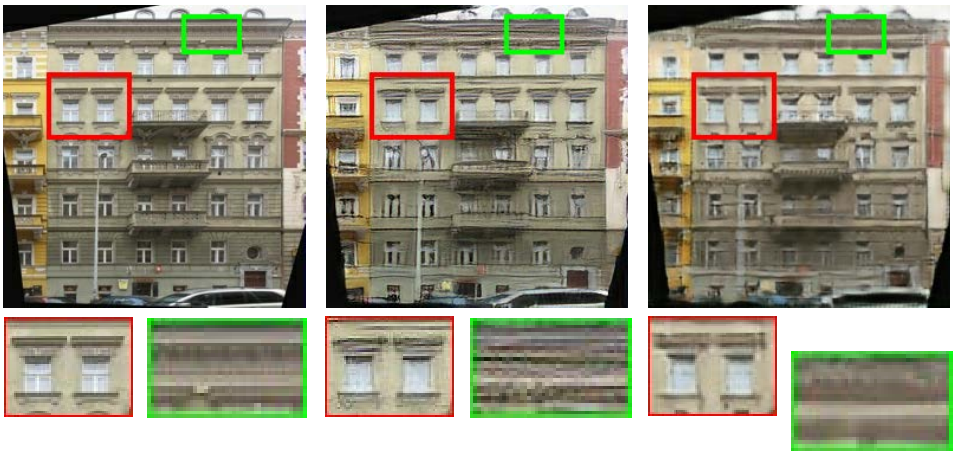


Fig. 10. Visual comparisons of our model with other methods for ×4 scale factor. From left to right: the original HR image, bicubic interpolation, SRCNN [15], LapSRN [35], our proposed Stage-GAN. Corresponding PSNR(dB) [42] and SSIM [43] are shown in bottom. Our method provides much cleaner and sharper results, whereas other methods produce blurry boundary.





Ground Truth PSNR/SSIM

SRGAN 21.04/0.5016

Ours 21.62/0.5376

Fig. 11. Visual comparisons of our model with SRGAN [33] for ×4 scale factor. Our results have fine details, such as the cornice of the building, whereas the SRGAN can not give good details in visual quality.

ing instances, such as facade, window, cornice, sill and so on, we use the low-frequency information of LR image and corresponding semantic information to reconstruct HR image. As shown in **Fig. 6**, by utilizing the semantic information for jointly reconstruction, the Stage-1 GAN with semantic maps can recover more details accurately compared to the Stage-1 GAN without semantic maps. As shown in **Fig. 7**, the reconstructed images with semantic maps are fed to Stage-2 GAN, yielding much cleaner images than the reconstructed images without semantic maps.

Fig. 8 illustrates some examples of Stage-0, Stage-1 and Stage-2 images generated by our Stage-GAN with scale factor ×4. As shown in **Fig. 8(e)**, Stage-1 GAN fails to produce HR images with sharp edges. In order to improve the quality of the images, we propose the Stage-2 GAN as a post-processing module to refine the results of Stage-1 GAN, as shown in **Fig. 3**. In the last row **Fig. 8(f)**, Stage-2 GAN can generate HR images with photo-realistic details. For scale factor ×8, as shown in **Fig. 9**, the images generated by Stage-2 GAN have much cleaner and sharper details than the results of Stage-1 GAN, which validates the importance of Stage-2 GAN for image SR.

3.3 Comparisons with Other Methods

To validate our method, we provide quantitative and qualitative comparisons with several SR methods in scaling factors ×4 and ×8, which include bicubic [6], SRCNN [15], LapSRN [35] and SRGAN [33]. In this work, we do not achieve better performance on PSNR and SSIM, but instead generate photo-realistic HR images with high perceptual quality. In particular, for ×8 SR, our method achieves higher GSM [44] values, which has been correlated well with human perception.

In **Fig. 10**, we show visual comparisons of our Stage-GAN with other methods for ×4 super-resolution. Compared to the other methods, our model trained for feature reconstruction does a very good job at reconstructing sharpen edges and fine details, such as the windows in the above images. We observe that the other methods reconstruct results with noticeable blurring. In contrast, our approach effectively suppresses such blurring through the models of Stage-GAN and the robust loss function. Our model can generate the sharper edges of buildings and achieve a good performance in the visual results, but compare to the original HR image, the reconstructed image may have little difference in structure or the shift of corresponding pixel-to-pixel, which harms its PSNR and SSIM compared to baseline methods.

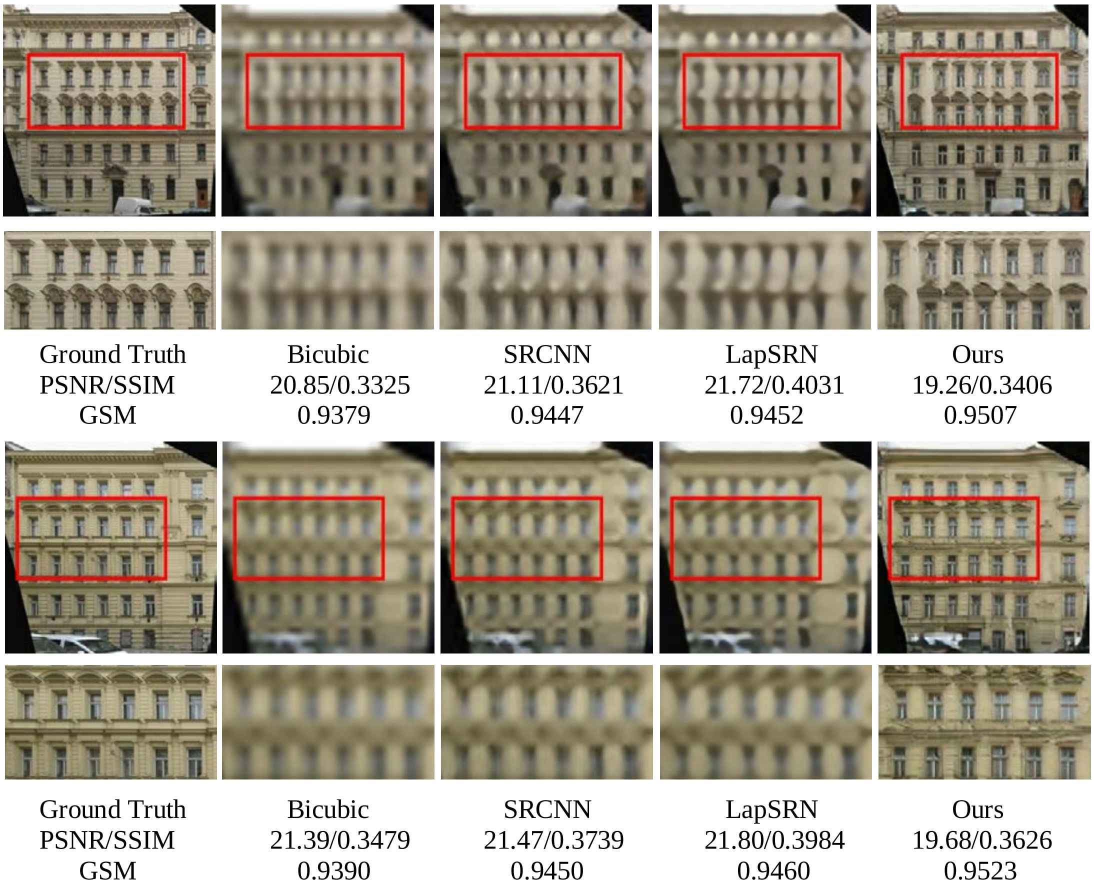


Fig. 12. Visual comparisons of our model with other methods for ×8 scale factor. From left to right: the original HR image, bicubic interpolation, SRCNN [15], LapSRN [35], our proposed Stage-GAN. Corresponding PSNR(dB) [42], SSIM [43] and GSM [44] are shown in bottom. Our method provides much cleaner and sharper results, whereas other methods produce blurry boundary.

In **Fig. 11**, we compare our model to SRGAN [33] on several images with the scaling factor ×4. Our method achieves better performance on PSNR and SSIM than SRGAN [33]. Due to the combination of low-frequency information and semantic information, our results have fine details, such as the cornice and facade of buildings, whereas SRGAN can not give good details in visual quality.

For ×8 SR, it is challenging to predict HR images from LR images. We show visual comparisons on Facades dataset with ×8 scale factor in **Fig. 12**. It can be observed from the visual results that our Stage-GAN can produce much more visually pleasant HR images than the comparing SR methods. Specifically, one can see that the performance of SRCNN [15] and LapSRN [35] is severely affected by the LR images, due to limited features available in the LR spaces. From the top images in **Fig. 12**, we can see that SRCNN [15] and LapSRN [35] both tend to produce over-smoothed results, whereas our Stage-GAN can recover sharp images with better intensity and gradient statistics of clean images. Our method achieves better performance on GSM [44], which emphasis on gradient similarity and measures the change in structure in images. We see that our approach does a good performance at edges and fine details compared to other methods, such as the windows, sill and cornice of buildings. The baseline methods do not super-resolve the fine structures well. In contrast, our method reconstructs high-quality HR images with photo-realistic details.

4. Conclusion

In this paper, we propose Stage Generative Adversarial Networks (Stage-GAN) with semantic maps for image SR in large scaling factors (×4, ×8). We decompose the task of image SR into a novel semantic map based reconstruction and refinement process. The Stage-0 GAN generates semantic maps from given LR images. The Stage-1 GAN reconstructs the high-resolution images by conditioning on the semantic maps and corresponding LR images. The Stage-2 GAN refines the Stage-1 results and sharpens the edges of objects, yielding high-resolution images with more photo-realistic details. Extensive experiments and comparisons with other SR methods demonstrate the effectiveness of our proposed method, and our method performs favorably against the compared SR methods in terms of visual quality.

References

[1] S. Peled, and Y. Yeshurun, “Superresolution in MRI: Application to human white matter fiber tract visualization by diffusion tensor imaging,” *Magnetic Resonance in Medicine Official Journal of the Society of Magnetic Resonance in Medicine*, vol.45, no.1, pp. 29-35, Jan. 2001. Article (CrossRef Link).

[2] W. Shi, J. Caballero, C. Ledig, X. Zhuang, W. Bai, K. Bhatia, A. Marvao, T. Dawes, D. ORegan, and D. Rueckert, “Cardiac Image Super-Resolution with Global Correspondence using Multi-Atlas PatchMatch,” *Medical Image Computing and Computer Assisted Intervention*, pp. 9-16, 2013. Article (CrossRef Link).

[3] L. Zhang, H. Zhang, H. Shen, and P. Li, “A super-resolution reconstruction algorithm for surveillance images,” *IEEE Transactions on Signal Processing*, vol. 90, no. 3, pp. 848-859, Sep. 2009. Article (CrossRef Link).

[4] M. W. Thornton, P. M. Atkinson, and D. A. Holland, “Sub-pixel mapping of rural land cover objects from fine spatial resolution satellite sensor imagery using super-resolution pixel-swapping,” *International Journal of Remote Sensing*, vol. 27, no. 3, pp. 473–491, Feb. 2007. Article (CrossRef Link).

[5] B. K. Gunturk, A. U. Batur, Y. Altunbasak, M. H. Hayes, and R. M. Mersereau, “Eigenface-domain super-resolution for face recognition,” *IEEE Transactions on Image Processing*, vol.12, no. 5, pp. 597-606, May 2003. Article (CrossRef Link).

[6] R. Keys, “Cubic convolution interpolation for digital image processing,” *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 29, no.6, pp. 1153-1160, Dec. 1981. Article (CrossRef Link).

[7] K. In Kim, and Y. Kwon, “Single-Image Super-Resolution Using Sparse Regression and Natural Image Prior,” *IEEE Transactions on Pattern Analysis & Machine Intelligence*, vol.32, no. 6, pp. 1127-1133, June 2010. Article (CrossRef Link).

[8] J. Sun, Z. Xu, and H.-Y. Shum, “Image super-resolution using gradient profile prior,” in *Proc. of IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1-8, 2008. Article (CrossRef Link).

[9] D. Glasner, S. Bagon and M. Irani, "Super-resolution from a single image," in *Proc. of IEEE International Conference on Computer Vision*, pp.349-356, 2009. Article (CrossRef Link).

[10] J.-B. Huang, A. Singh, and N. Ahuja, "Single image super-resolution from transformed self-exemplars," in *Proc. of IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5197-5206, 2015. Article (CrossRef Link).

[11] F. Wang, M. Jiang, C. Qian, S. Yang, C. Li, H. Zhang, X. Wang and X. Tang, "Residual attention network for image classification," in *Proc. of IEEE Conference on Computer Vision and Pattern Recognition*, pp.6450-6458, 2017. Article (CrossRef Link).

[12] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-Image Translation with Conditional Adversarial Networks," in *Proc. of IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5967-5976, 2017. Article (CrossRef Link).

[13] Z. Li, J. Tang, and T. Mei, "Deep Collaborative Embedding for Social Image Understanding," *IEEE Transactions on Pattern Analysis & Machine Intelligence*, vol. 41, no. 9, pp. 2070-2083, 2019. Article (CrossRef Link).

[14] Z. Li, and J. Tang, "Weakly supervised deep matrix factorization for social image understanding," *IEEE Transactions on Image Processing*, vol. 26, no. 1, pp. 276-288, Jan. 2017. Article (CrossRef Link).

[15] C. Dong, C. C. Loy, K. He, and X. Tang, "Image Super-Resolution Using Deep Convolutional Networks," *IEEE Transactions on Pattern Analysis & Machine Intelligence*, vol. 38, no. 2, pp.295-307, Feb. 2016. Article (CrossRef Link).

[16] J. Kim, JK. Lee, and KM. Lee, "Deeply-Recursive Convolutional Network for Image Super-Resolution," in *Proc. of IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1637-1645, 2016. Article (CrossRef Link).

[17] J. Pan, S. Liu, J. Zhang, Y. Liu, J. Ren, Z. Li, J. Tang, H. Lu, Y.-W. Tai, and Ming-Hsuan Yang, "Learning Dual Convolutional Neural Networks for Low-Level Vision," in *Proc. of IEEE Conference on Computer Vision and Pattern Recognition*, pp.3070-3079, 2018. Article (CrossRef Link).

[18] B. Lim, S. Son, H. Kim, S. Nah, and K. M. Lee, "Enhanced deep residual networks for image super-resolution," in *Proc. of IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2017. Article (CrossRef Link).

[19] Y. Zhang, Y. Tian, Y. Kong, B. Zhong, and Y. Fu, "Residual dense network for image super-resolution," in *Proc. of IEEE Conference on Computer Vision and Pattern Recognition*, pp.2474-2481, 2018. Article (CrossRef Link).

[20] J. Bruna, P. Sprechmann, and Y. LeCun, "Super-Resolution with Deep Convolutional Sufficient Statistics," in *Proc. of IEEE International Conference on Learning Representations*, 2016. Article (CrossRef Link).

[21] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, and D. Warde-Farley, S. Ozair, A. Courville and Y. Bengio, "Generative adversarial nets," in *Proc. of International Conference on Neural Information Processing Systems MIT Press*, pp. 2672-2680, 2014. Article (CrossRef Link).

[22] M. Arjovsky, S. Chintala, and L. Bottou, "Wasserstein GAN," in *Proc. of International Conference on Machine Learning*, 2017. Article (CrossRef Link).

[23] J. Zhao, M. Mathieu, and Y. LeCun, "Energy-based Generative Adversarial Network," in *Proc. of International Conference on Learning Representations*, pp. 1-17, 2017. Article (CrossRef Link).

[24] J. Y. Zhu, P. Krahenbuhl, E. Shechtman, and A. A. Efros, "Generative Visual Manipulation on the Natural Image Manifold," in *Proc. of European Conference on Computer Vision*, pp. 597-613, 2016. Article (CrossRef Link).

[25] M. Mathieu, J. Zhao, A. Ramesh, P. Sprechmann, and Y. LeCun, "Disentangling factors of variation in deep representations using adversarial training," in *Proc. of International Conference on Neural Information Processing Systems*, pp. 5040-5048, 2016. Article (CrossRef Link).

[26] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen, "Improved Techniques for Training GANs," in *Proc. of International Conference on Neural Information Processing Systems*, 2016. Article (CrossRef Link).

[27] S. Reed, Z. Akata, X. Yan, L. Logeswaran, B. Schiele, and H. Lee, “Generative adversarial text to image synthesis,” in *Proc. of International Conference on International Conference on Machine Learning*, pp. 1060-1069, 2016. Article (CrossRef Link).

[28] D. Emily, C. Soumith, S. Arthur, and F. Rob, “Deep Generative Image Models using a Laplacian Pyramid of Adversarial Networks,” in *Proc. of International Conference on Neural Information Processing Systems*, pp. 1486-1494, 2015. Article (CrossRef Link).

[29] M. Mirza, and S. Osindero, “Conditional generative adversarial nets,” *arXiv preprint arXiv: 1411.1784*, 2014. Article (CrossRef Link).

[30] D. Pathak, P. Krahenbuhl, J. Donahue, T. Darrell, and A. A. Efros, “Context Encoders: Feature Learning by Inpainting,” in *Proc. of IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2536-2544, 2016. Article (CrossRef Link).

[31] H. Zhang, T. Xu, H. Li, S. Zhang, X. Wang, X. Huang, and D. Metaxas, “StackGAN: Text to Photo-Realistic Image Synthesis with Stacked Generative Adversarial Networks,” in *Proc. of IEEE International Conference on Computer Vision*, pp.5908-5916, 2017. Article (CrossRef Link).

[32] C. Li, and M. Wand, “Precomputed Real-Time Texture Synthesis with Markovian Generative Adversarial Networks,” in *Proc. of European Conference on Computer Vision*, pp.702-716, 2016. Article (CrossRef Link).

[33] C. Ledig, L. Theis, F. Husz´ar, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, Z. Wang, and W. Shi, “Photo-Realistic Single Image Super-Resolution Using a Generative Adversarial Network,” in *Proc. of IEEE Conference on Computer Vision and Pattern Recognition*, pp. 105-114, 2017. Article (CrossRef Link).

[34] A. Bulat, J. Yang, and G. Tzimiropoulos, “To learn image super-resolution, use a GAN to learn how to do image degradation first,” in *Proc. of European Conference on Computer Vision*, pp. 187-202, 2018. Article (CrossRef Link).

[35] W.-S. Lai, J.-B. Huang, N. Ahuja, and M.-H. Yang, “Deep Laplacian Pyramid Networks for Fast and Accurate Super-Resolution,” in *Proc. of IEEE Conference on Computer Vision and Pattern Recognition*, pp.5835-5843, 2017. Article (CrossRef Link).

[36] R. Olaf, F. Philipp, and B. Thomas, “U-Net: Convolutional Networks for Biomedical Image Segmentation,” in *Proc. of the International Conference on Medical Image Computing and Computer Assisted Intervention*, pp. 234-241, 2015. Article (CrossRef Link).

[37] J. Long, E. Shelhamer, and T. Darrell, “Fully convolutional networks for semantic segmentation,” in *Proc. of IEEE Conference on Computer Vision and Pattern Recognition*, pp.3431-3440, 2015. Article (CrossRef Link).

[38] S. Ioffe, and C. Szegedy, “Batch normalization: accelerating deep network training by reducing internal covariate shift,” in *Proc. of International Conference on International Conference on Machine Learning*, pp.448-456, 2015. Article (CrossRef Link).

[39] K. He, X. Zhang, S. Ren, and J. Sun, “Deep Residual Learning for Image Recognition,” in *Proc. of IEEE Conference on Computer Vision and Pattern Recognition*, pp.770-778, 2016. Article (CrossRef Link).

[40] R. Tyleček, and R. Šára, “Spatial Pattern Templates for Recognition of Objects with Regular Structure,” in *Proc. of German Conference Pattern Recognition*, pp.364-374, 2013. Article (CrossRef Link).

[41] D. Kingma, and J. Ba, “Adam: A method for Stochastic Optimization,” *arXiv preprint arXiv: 1412.6980*, 2014. Article (CrossRef Link).

[42] M. Irani, and S. Peleg, “Motion Analysis for Image Enhancement: Resolution, Occlusion, and Transparency,” *Journal of Visual Communications and Image Representation*, vol. 4, no.4, pp. 324-335, Dec. 1993. Article (CrossRef Link).

[43] Z. Wang, A.C. Bovik, H. R. Sheikh, and E. P. Simoncelli, “Image quality assessment: from error visibility to structural similarity,” *IEEE Transactions on Image Processing*, vol.13, no.4, pp.600-612, April 2004. Article (CrossRef Link).

[44] A. Liu, W. Lin, and M. Narwaria, "Image Quality Assessment Based on Gradient Similarity," *IEEE Transactions on Image Processing*, vol.21, no.4, pp.1500-1512, April 2012. Article (CrossRef Link).

Zhensong Wei received the B.S. degree in Electronic Information Science and Technology in 2017 from the School of Information Science and Engineering, Henan University of Technology. He is currently pursuing the M.S. degree in the Institute of Information Science, Beijing Jiaotong University. He works in image super-resolution and image compression.

Huihui Bai received her B.S. degree from Beijing Jiaotong University, China, in 2001, and her Ph.D. degree from Beijing Jiaotong University, China, in 2008. She is currently a professor in Beijing Jiaotong University. She has been engaged in R&D work in video coding technologies and standards, such as HEVC, 3D video compression, multiple description video coding (MDC), and distributed video coding (DVC).

Yao Zhao received the B.S. degree from Fuzhou University, China, in 1989, and the ME degree from Southeast University, Nanjing, China, in 1992, both from the Radio Engineering Department, and the PhD degree from the Institute of Information Science, Beijing Jiaotong University (BJTU), China, in 1996. He became an associate professor at BJTU in 1998 and became a professor in 2001. From 2001 to 2002, he was a senior research fellow with the Information and Communication Theory Group, Faculty of Information Technology and Systems, Delft University of Technology, Delft, The Netherlands. He is currently the director of the Institute of Information Science, BJTU. His current research interests include image/video coding, digital watermarking and forensics, and video analysis and understanding. He serves on the editorial boards of several international journals, including as associate editors of IEEE Transactions on Cybernetics, IEEE Signal Processing Letters, and an area editor of Signal Processing: Image Communication (Elsevier), etc. He was named a distinguished young scholar by the National Science Foundation of China in 2010, and was elected as a Chang Jiang Scholar of Ministry of Education of China in 2013. He is a senior member of the IEEE.