# Structure-Preserving Super Resolution with Gradient Guidance

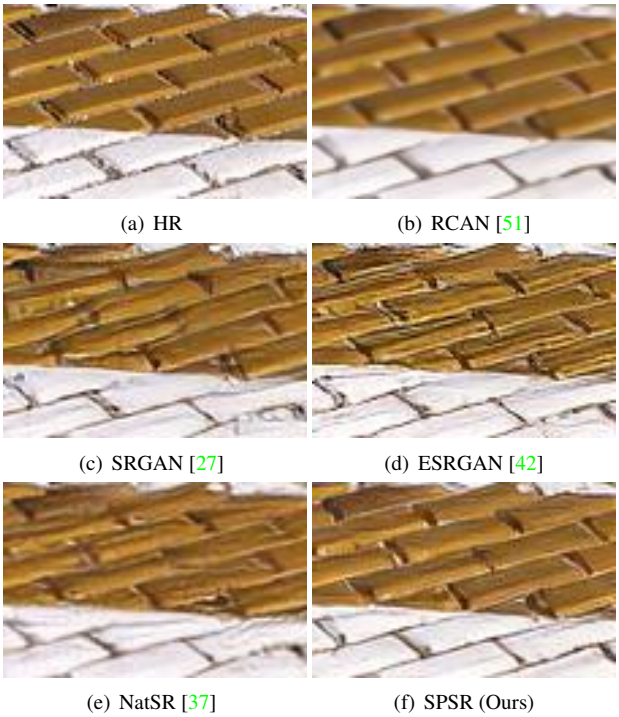
Cheng Ma[1,2,3], Yongming Rao[1,2,3], Yean Cheng[1], Ce Chen[1], Jiwen Lu[1,2,3*], Jie Zhou[1,2,3,4]
[1]Department of Automation, Tsinghua University, China
[2]State Key Lab of Intelligent Technologies and Systems, China
[3]Beijing National Research Center for Information Science and Technology, China
[4]Tsinghua Shenzhen International Graduate School, Tsinghua University, China
macheng17@mails.tsinghua.edu.cn; raoyongming95@gmail.com
{cya17, chence17}@mails.tsinghua.edu.cn; {lujiwen, jzhou}@tsinghua.edu.cn

## Abstract

*Structures matter in single image super resolution (SISR). Recent studies benefiting from generative adversarial network (GAN) have promoted the development of SISR by recovering photo-realistic images. However, there are always undesired structural distortions in the recovered images. In this paper, we propose a structure-preserving super resolution method to alleviate the above issue while maintaining the merits of GAN-based methods to generate perceptual-pleasant details. Specifically, we exploit gradient maps of images to guide the recovery in two aspects. On the one hand, we restore high-resolution gradient maps by a gradient branch to provide additional structure priors for the SR process. On the other hand, we propose a gradient loss which imposes a second-order restriction on the super-resolved images. Along with the previous image-space loss functions, the gradient-space objectives help generative networks concentrate more on geometric structures. Moreover, our method is model-agnostic, which can be potentially used for off-the-shelf SR networks. Experimental results show that we achieve the best PI and LPIPS performance and meanwhile comparable PSNR and SSIM compared with state-of-the-art perceptual-driven SR methods. Visual results demonstrate our superiority in restoring structures while generating natural SR images.* [1]

## 1. Introduction

Single image super resolution (SISR) aims to recover high-resolution (HR) images from their low-resolution (LR) counterparts. SISR is a fundamental problem in the community of computer vision and can be applied in many image analysis tasks including surveillance and satellite image. It is a widely known ill-posed problem since each LR input may have multiple HR solutions. With the development of deep learning, a number of SR methods [8, 35] have been proposed. Most of them are optimized by the mean squared error (MSE) which measures the pixel-wise distances between SR images and the HR ones. However, such opti-



(a) HR (b) RCAN [51] (c) SRGAN [27] (d) ESRGAN [42] (e) NatSR [37] (f) SPSR (Ours)

Figure 1. SR results of different methods. RCAN represents PSNR-oriented methods, typically generating straight but blurry edges for the bricks. Perceptual-driven methods including SRGAN, ESRGAN and NatSR commonly recover sharper but geometric-inconsistent textures. Our SPSR result is sharper than that of RCAN, and preserve finer geometric structures compared with perceptual-driven methods. Best viewed on screen.

*Corresponding author
[1]Code: https://github.com/Maclory/SPSR

mizing objective impels a deep model to produce an image which may be a statistical average of possible HR solutions to the one-to-many problem. As a result, such methods usually generate blurry images with high peak signal-to-noise ratio (PSNR).

Hence, several methods aiming to recover photo-realistic images have recently utilized the generative adversarial network (GAN) [15], such as SRGAN [27], EnhanceNet [34], ESRGAN [42] and NatSR [37]. While GAN-based methods can generate high-fidelity SR results, there are always geometric distortions along with sharp edges and fine textures. Some SR examples are presented in Figure 1. We can see RCAN [51] recovers blurry but straight edges for the bricks, while edges restored by perceptual-driven methods are sharper but twisted. In fact, GAN-based methods generally suffer from structural inconsistency since the discriminators may introduce unstable factors to the optimization procedure. Some methods have been proposed to balance the trade-off between the merits of two kinds of SR methods. For example, Controllable Feature Space Network (CFSNet) [40] designs an interactive framework to transfer continuously between two objectives of perceptual quality and distortion reduction. Nevertheless, the intrinsic problem is not mitigated since the two goals cannot be achieved simultaneously. Hence it is necessary to explicitly guide perceptual-driven SR methods to preserve structures for further enhancing the SR performance.

In this paper, we propose a structure-preserving super resolution method to alleviate the above-mentioned issue. Since the gradient map reveals the sharpness of each local region in an image, we exploit this powerful tool to guide image recovery. On the one hand, we design a gradient branch which converts the gradient maps of LR images to the HR ones as an auxiliary SR problem. The recovered gradients can be integrated into the SR branch to provide structure prior for SR. Besides, the gradients can highlight the regions where sharpness and structures should be paid more attention to, so as to guide the high-quality generation explicitly. This idea is motivated by the observation that once edges are recovered with high-fidelity, the SR task can be treated as a color-filling problem with strong clues given by the LR images. On the other hand, we propose a gradient loss to explicitly supervise the gradient maps of recovered images. Together with the image-space loss functions in existing methods, the gradient loss restricts the second-order relationship of neighboring pixels. Hence the structural configuration can be better retained with such guidance, and the SR results with high perceptual quality and fewer geometric distortions can be obtained. Moreover, our method is model-agnostic, which can be potentially used for off-the-shelf SR networks. To the best of our knowledge, we are the first to explicitly consider preserving geometric structures in GAN-based SR methods. Experimental results on benchmark datasets show that our method succeeds in enhancing SR fidelity by reducing structural distortions.

## 2. Related Work

Here we review SISR methods [7, 10, 12, 13, 14, 19, 22, 25, 38, 44, 46, 47] which can be classified into two categories: PSNR-oriented methods and perceptual-driven ones. We also investigate methods relevant to gradient.

**PSNR-Oriented Methods**: Most previous approaches target high PSNR. As a pioneer, Dong *et al*. [8] propose SRCNN, which firstly maps LR images to HR ones by a three-layer CNN. DRCN [24] and VDSR [23] are further proposed by Kim *et al*. to improve SR performance. Moreover, Ledig *et al*. [27] propose SRResNet by employing the idea of ResNet [17]. Zhang *et al*. [52] propose RDN by utilizing residual dense blocks in the SR framework. They further introduce RCAN [51] and achieve superior performance on PSNR. Li *et al*. [28] propose a feedback framework to refine the super-resolved results step by step.

**Perceptual-Driven Methods**: The methods mentioned above all focus on achieving high PSNR and thus use the MSE loss or L1 loss as loss functions. However, these methods usually produce blurry images. Johnson *et al*. [20] propose perceptual loss to improve the visual quality of recovered images. Ledig *et al*. [27] utilize adversarial loss [15] to construct SRGAN, which becomes the first framework able to generate photo-realistic HR images. Furthermore, Sajjadi *et al*. [34] restore high-fidelity textures by texture loss. Wang *et al*. [42] enhance the previous frameworks by introducing Residual-in-Residual Dense Block (RRDB) to the proposed ESRGAN. Wang *et al*. [41] exploit semantic segmentation maps as priors to generate more natural textures for specific categories. Rad *et al*. [32] propose a targeted perceptual loss on the basis of the labels of object, background and boundary. Although these existing perceptual-driven methods indeed improve the overall visual quality of super-resolved images, they sometimes generate unnatural artifacts including geometric distortions when recovering details.

**Gradient-Relevant Methods**: Gradient information has been utilized in previous work [2, 29]. For SR methods, Fattal [11] proposes a method based on edge statistics of image gradients by learning the prior dependency of different resolutions. Sun *et al*. [39] propose a gradient profile prior to represent image gradients and a gradient field transformation to enhance sharpness of super-resolved images. Yan *et al*. [45] propose a SR method based on gradient profile sharpness which is extracted from gradient description models. In these methods, statistical dependencies are modeled by estimating HR edge-related parameters according to those observed in LR images. However, the modeling procedure is accomplished point by point, which is complex and inflexible. In fact, deep learning is outstand-

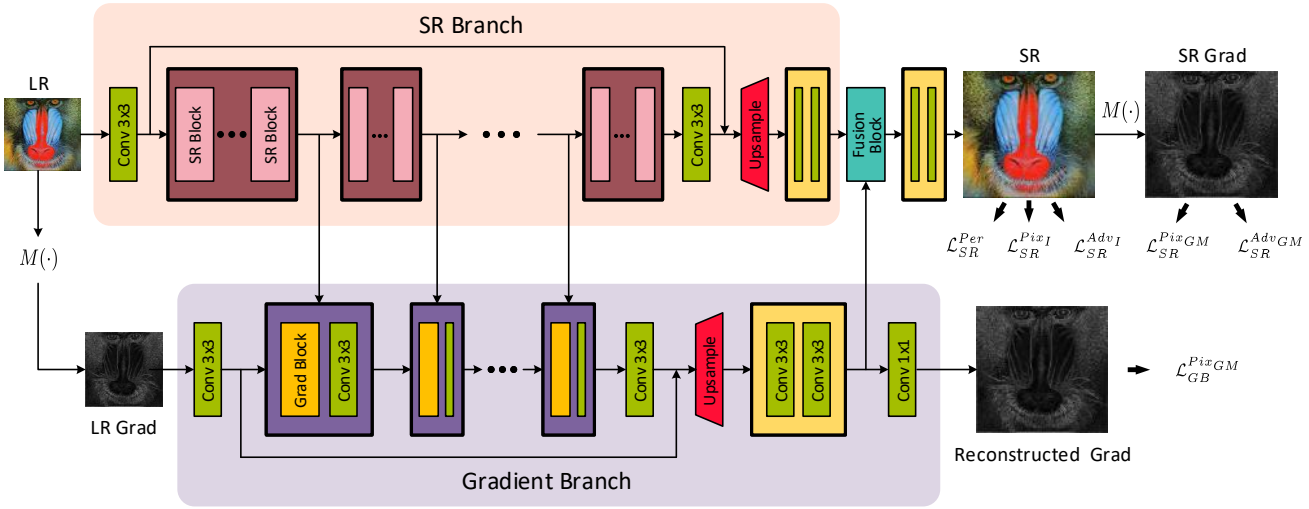



Figure 2. Overall framework of our SPSR method. Our architecture consists of two branches, the SR branch and the gradient branch. The gradient branch aims to super-resolve LR gradient maps to the HR counterparts. It incorporates multi-level representations from the SR branch to reduce parameters and outputs gradient information to guide the SR process by a fusion block in turn. The final SR outputs are optimized by not only conventional image-space losses, but also the proposed gradient-space objectives.

ing in handling probability transformation over the distribution of pixels. However, few methods have utilized its powerful abilities in gradient-relevant SR methods. Moreover, Zhu *et al*. [53] propose a gradient-based SR method by collecting a dictionary of gradient patterns and modeling deformable gradient compositions. Yang *et al*. [48] propose a recurrent residual network to reconstruct fine details guided by the edges which are extracted by off-the-shelf edge detector. While edge reconstruction and gradient field constraint have been utilized in some methods, their purposes are mainly to recover high-frequency components for PSNR-orientated SR methods. Different from these methods, we aim to reduce geometric distortions produced by GAN-based methods and exploit gradient maps as structure guidance for SR. For deep adversarial networks, gradient-space constraint may provide additional supervision for better image reconstruction. To the best of our knowledge, no GAN-based SR method has exploited gradient-space guidance for preserving texture structures. In this work, we aim to leverage gradient information to further improve the GAN-based SR methods.

# 3. Approach

In this section, we first introduce the overall framework. Then we present the details of gradient branch, attentive fusion module and final objective functions accordingly.

## 3.1. Overview

In SISR, we aim to take LR images $I^{LR}$ as inputs and generate SR images $I^{SR}$ given their HR counterparts $I^{HR}$ as ground-truth. We denote the generator as $G$ and its parameters as $\theta_G$ and then we have $I^{SR} = G(I^{LR}; \theta_G)$. $I^{SR}$ should be as similar to $I^{HR}$ as possible. If the parameters are optimized by an loss function $\mathcal{L}$, we have the following formulation:

$$\theta_G^* = \arg\min_{\theta_G} \mathbb{E}_{I^{SR}} \mathcal{L}(G(I^{LR}; \theta_G), I^{HR}). \quad (1)$$

The overall framework is depicted as Figure 2. The generator is composed of two branches, one of which is a structure-preserving SR branch and the other is a gradient branch. The SR branch takes $I^{LR}$ as input and aims to recover the SR output $I^{SR}$ with the guidance provided by the SR gradient map from the gradient branch.

## 3.2. Details in Architecture

### 3.2.1 Gradient Branch

The target of the gradient branch is to estimate the translation of gradient maps from the LR modality to the HR one. The gradient map for an image $I$ is obtained by computing the difference between adjacent pixels:

$$\begin{aligned} I_x(\mathbf{x}) &= I(x+1, y) - I(x-1, y), \\ I_y(\mathbf{x}) &= I(x, y+1) - I(x, y-1), \\ \nabla I(\mathbf{x}) &= (I_x(\mathbf{x}), I_y(\mathbf{x})), \\ M(I) &= \|\nabla I\|_2, \end{aligned} \quad (2)$$

where $M(\cdot)$ stands for the operation to extract gradient map whose elements are gradient lengths for pixels with coordinates $\mathbf{x} = (x, y)$. The operation to get the gradients can be easily achieved by a convolution layer with a fixed

kernel. In fact, we do not consider gradient direction information since gradient intensity is adequate to reveal the sharpness of local regions in recovered images. Hence we adopt the intensity maps as the gradient maps. Such gradient maps can be regarded as another kind of images, so that techniques for image-to-image translation can be utilized to learn the mapping between two modalities. The translation process is equivalent to the spatial distribution translation from LR edge sharpness to HR edge sharpness. Since most area of the gradient map is close to zero, the convolutional neural network can concentrates more on the spatial relationship of outlines. Therefore, it may be easier for the network to capture structure dependency and consequently produce approximate gradient maps for SR images.

As shown in Figure 2, the gradient branch incorporates several intermediate-level representations from the SR branch. The motivation of such scheme is that the well-designed SR branch is capable of carrying rich structural information which is pivotal to the recovery of gradient maps. Hence we utilize the features as a strong prior to promote the performance of the gradient branch, whose parameters can be largely reduced in this case. Between each two intermediate features, there is a gradient block which can be any basic block to extract higher-level features. Once we get the SR gradient maps by the gradient branch, we are able to integrate the obtained gradient features into the SR branch to guide SR reconstruction in turn. The magnitude of gradient map can implicitly reflect whether a recovered region should be sharp or smooth. In practice, we feed the feature maps produced by the next-to-last layer of gradient branch to the SR branch. Meanwhile, we generate the output gradient maps by a $1 \times 1$ convolution layer with these feature maps as inputs.

#### 3.2.2 Structure-Preserving SR Branch

We design a structure-preserving SR branch to get the final SR outputs. This branch constitutes of two parts. The first part is a regular SR network comprising of multiple generative neural blocks which can be any architecture. Here we introduce the Residual in Residual Dense Block (RRDB) proposed in ESRGAN [42]. There are 23 RRDB blocks in the original model. Therefore, we incorporate the feature maps from the 5th, 10th, 15th, 20th blocks to the gradient branch. Since regular SR models produce images with only 3 channels, we remove the last convolutional reconstruction layer and feed the output feature to the consecutive part. The second part of the SR branch wires the SR gradient feature maps obtained from the gradient branch as mentioned above. We fuse the structure information by a fusion block which fuses the features from two branches together. Specifically, we concatenate the two features and then use another RRDB block and convolutional layer to reconstruct the final SR features. It is noteworthy that we only add one RRDB block into the SR branch. Thus the parameter increment is slight compared to the original model with 23 blocks.

### 3.3. Objective Functions

**Conventional Loss**: Most SR methods optimize the elaborately designed networks by a common pixelwise loss, which is efficient for the task of super resolution measured by PSNR. This metric can reduce the average pixel difference between recovered images and ground-truths but the results may be too smooth to maintain sharp edges for visual effects. However, this loss is still widely used to accelerate convergence and improve SR performance:

$$\mathcal{L}_{SR}^{Pix_I} = \mathbb{E}_{I^{SR}} \|G(I^{LR}) - I^{HR}\|_1. \quad (3)$$

Perceptual loss has been proposed in [20] to improve perceptual quality of recovered images. Features containing semantic information are extracted by a pre-trained VGG network [36]. The Euclidean distances between the features of HR images and SR ones are minimized in perceptual loss:

$$\mathcal{L}_{SR}^{Per} = \mathbb{E}_{I^{SR}} \|\phi_i(G(I^{LR})) - \phi_i(I^{HR})\|_1, \quad (4)$$

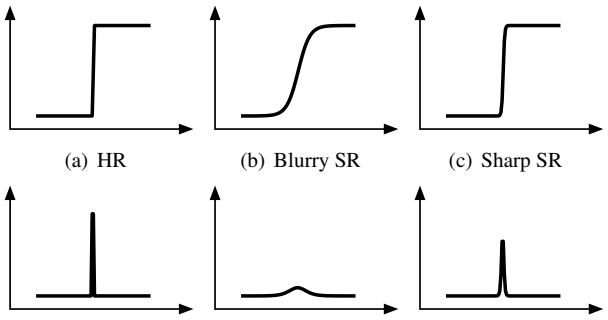
where $\phi_i(.)$ denotes the $i$th layer output of the VGG model.

Methods [27, 42] based on generative adversarial networks (GANs) [3, 4, 15, 16, 21, 33] also play an important role in the SR problem. The discriminator $D_I$ and the generator $G$ are optimized by a two-player game as follows:

$$\begin{aligned} \mathcal{L}_{SR}^{Dis_I} &= -\mathbb{E}_{I^{SR}}[\log(1 - D_I(I^{SR}))] \\ &\quad -\mathbb{E}_{I^{HR}}[\log D_I(I^{HR})], \quad (5) \\ \mathcal{L}_{SR}^{Adv_I} &= -\mathbb{E}_{I^{SR}}[\log D_I(G(I^{LR}))]. \quad (6) \end{aligned}$$

Following [21, 42] we conduct relativistic average GAN (RaGAN) to achieve better optimization in practice. Models supervised by the above objective functions merely consider the image-space constraint for images, but neglect the semantically structural information provided by the gradient space. While the generated results look photo-realistic, there are also a number of undesired geometric distortions. Thus we introduce the gradient loss to alleviate this issue.

**Gradient Loss**: Our motivation can be illustrated clearly by Figure 3. Here we only consider a simple 1-dimensional case. If the model is only optimized in image space by the L1 loss, we usually get a SR sequence as Figure 3 (b) given an input testing sequence whose ground-truth is a sharp edge as Figure 3 (a). The model fails to recover sharp edges for the reason that the model tends to give an statistical average of possible HR solutions from training data. In this case, if we compute and show the gradient magnitudes of two sequences, it can be observed that the SR gradient is flat with low values while the HR gradient is a spike with



(a) HR (b) Blurry SR (c) Sharp SR

(d) HR Gradiant (e) Blurry Gradiant (f) Sharp Gradiant

Figure 3. An illumination of a simple 1-D case. The first row shows the pixel sequences and the second row shows their corresponding gradient maps.

high values. They are far from each other. This inspires us that if we add a second-order gradient constraint to the optimization objective, the model may learn more from the gradient space. It helps the model focus on neighboring configuration, so that the local intensity of sharpness can be inferred more appropriately. Therefore, if the gradient information as Figure 3 (f) is captured, the probability of recovering Figure 3 (c) is increased significantly. SR methods can benefit from such guidance to avoid over-smooth or over-sharpening restoration. Moreover, it is easier to extract geometric characteristics in the gradient space. Hence geometric structures can be also preserved well, resulting in more photo-realistic SR images.

Here we propose a gradient loss to achieve the above goals. Since we have mentioned the gradient map is an ideal tool to reflect structural information of an image, it can also be utilized as a second-order constraint to provide supervision to the generator. We formulate the gradient loss by diminishing the distance between the gradient map extracted from the SR image and the one from the corresponding HR image. With the supervision in both image and gradient domains, the generator can not only learn fine appearance, but also attach importance to avoiding detailed geometric distortions. Therefore, we design two terms of loss to penalize the difference in the gradient maps (GM) of the SR and HR images. One is based on the pixelwise loss as follows:

$$\mathcal{L}_{SR}^{Pix_{GM}} = \mathbb{E}_{I^{SR}} \|M(G(I^{LR})) - M(I^{HR})\|_1. \quad (7)$$

The other is to discriminate whether a gradient patch is from the HR gradient map. We design another gradient discriminator network to achieve this goal:

$$\begin{aligned} \mathcal{L}_{SR}^{Dis_{GM}} &= -\mathbb{E}_{I^{SR}}[\log(1 - D_{GM}(M(I^{SR})))] \\ &\quad -\mathbb{E}_{I^{HR}}[\log D_{GM}(M(I^{HR}))]. \end{aligned} \quad (8)$$

The gradient discriminator can also supervise the generation of SR results by adversarial learning:

$$\mathcal{L}_{SR}^{Adv_{GM}} = -\mathbb{E}_{I^{SR}}[\log D_{GM}(M(G(I^{LR})))]. \quad (9)$$

Note that each step in the operation $M(\cdot)$ is differentiable. Hence the model with gradient loss can be trained in an end-to-end manner. Furthermore, it is convenient to adopt gradient loss as additional guidance in any generative model due to the concise formulation and strong transferability.

**Overall Objective**: In conclusion, we have two discriminators $D_I$ and $D_{GM}$ which are optimized by $\mathcal{L}_{SR}^{Dis_I}$ and $\mathcal{L}_{SR}^{Dis_{GM}}$, respectively. For the generator, two terms of loss are used to provide supervision signals simultaneously. One is imposed on the structure-preserving SR branch while the other is to reconstruct high-quality gradient maps by minimizing the pixelwise loss $\mathcal{L}_{GB}^{Pix_{GM}}$ in the gradient branch (GB). The overall objective is defined as follows:

$$\begin{aligned} \mathcal{L}^G &= \mathcal{L}_{SR}^G + \mathcal{L}_{GB}^G \\ &= \mathcal{L}_{SR}^{Per} + \beta_{SR}^I \mathcal{L}_{SR}^{Pix_I} + \gamma_{SR}^I \mathcal{L}_{SR}^{Adv_I} + \beta_{SR}^{GM} \mathcal{L}_{SR}^{Pix_{GM}} \\ &\quad + \gamma_{SR}^{GM} \mathcal{L}_{SR}^{Adv_{GM}} + \beta_{GB}^{GM} \mathcal{L}_{GB}^{Pix_{GM}}. \end{aligned} \quad (10)$$

$\beta_{SR}^I$, $\gamma_{SR}^I$, $\beta_{SR}^{GM}$, $\gamma_{SR}^{GM}$ and $\beta_{GB}^{GM}$ denote the trade-off parameters of different losses. Among these, $\beta_{SR}^I$, $\beta_{SR}^{GM}$ and $\beta_{GB}^{GM}$ are the weights of the pixel losses for SR images, gradient maps of SR images and SR gradient maps respectively. $\gamma_{SR}^I$ and $\gamma_{SR}^{GM}$ are the weights of the adversarial losses for SR image and their gradient maps.

# 4. Experiments

## 4.1. Implementation Details

**Datasets and Evaluation Metrics**: We evaluate the SR performance of our proposed SPSR method. We utilize DIV2K [1] as the training dataset and five commonly used benchmarks for testing: Set5 [5], Set14 [49], BSD100 [30], Urban100 [18] and General100 [9]. We downsample HR images by bicubic interpolation to get LR inputs and only consider the scaling factor of $4\times$ in our experiments. We choose Perceptual Index (PI) [6], Learned Perceptual Image Patch Similarity (LPIPS) [50], PSNR and Structure Similarity (SSIM) [43] as the evaluation metrics. Lower PI and LPIPS values indicate higher perceptual quality.

**Training Details**: We use the architecture of ESRGAN [42] as the backbone of our SR branch and the RRDB block [42] as the gradient block. We randomly sample 15 $32 \times 32$ patches from LR images for each input mini-batch. Therefore the ground-truth HR patches have a size of $128 \times 128$. We initialize the generator with the parameters of a pre-trained PSNR-oriented model. The pixelwise loss, perceptual loss, adversarial loss and gradient loss are used as the optimizing objectives. A pre-trained 19-layer VGG network [36] is employed to calculate the feature distances in the perceptual loss. We also use a VGG-style network to perform discrimination. ADAM optimizor [26] with $\beta_1 = 0.9, \beta_2 = 0.999$ and $\epsilon = 1 \times 10^{-8}$ is used for

Table 1. Comparison with state-of-the-art perceptual-driven SR methods on benchmark datasets. The best performance is **highlighted** in **red** (1st best) and **blue** (2nd best). Our SPSR obtains the best PI and LPIPS values and comparable PSNR and SSIM values simultaneously. NatSR is more like a PSNR-oriented method since it has high PSNR and SSIM and relatively poor PI and LPIPS performance.

| Dataset | Metric | Bicubic | SFTGAN [41] | SRGAN [27] | ESRGAN [42] | NatSR [37] | SPSR |
|---|---|---|---|---|---|---|---|
| **Set5** | PI | 7.3699 | 3.7587 | 3.9820 | **3.7522** | 4.1648 | **3.2743** |
| | LPIPS | 0.3407 | 0.0890 | 0.0882 | **0.0748** | 0.0939 | **0.0644** |
| | PSNR | 28.420 | 29.932 | 29.168 | **30.454** | **30.991** | 30.400 |
| | SSIM | 0.8245 | 0.8665 | 0.8613 | **0.8677** | **0.8800** | 0.8627 |
| **Set14** | PI | 7.0268 | **2.9063** | 3.0851 | 2.9261 | 3.1094 | **2.9036** |
| | LPIPS | 0.4393 | 0.1481 | 0.1663 | **0.1329** | 0.1758 | **0.1318** |
| | PSNR | 26.100 | 26.223 | 26.171 | 26.276 | **27.514** | **26.640** |
| | SSIM | 0.7850 | 0.7854 | 0.7841 | 0.7783 | **0.8140** | **0.7930** |
| **BSD100** | PI | 7.0026 | **2.3774** | 2.5459 | 2.4793 | 2.7801 | **2.3510** |
| | LPIPS | 0.5249 | 0.1769 | 0.1980 | **0.1614** | 0.2114 | **0.1611** |
| | PSNR | 25.961 | 25.505 | 25.459 | 25.317 | **26.445** | **25.505** |
| | SSIM | 0.6675 | 0.6549 | 0.6485 | 0.6506 | **0.6831** | **0.6576** |
| **General100** | PI | 7.9365 | **4.2878** | 4.3757 | 4.3234 | 4.6262 | **4.0991** |
| | LPIPS | 0.3528 | 0.1030 | 0.1055 | **0.0879** | 0.1117 | **0.0863** |
| | PSNR | 28.018 | 29.026 | 28.575 | 29.412 | **30.346** | **29.414** |
| | SSIM | 0.8282 | 0.8508 | 0.8541 | **0.8546** | **0.8721** | 0.8537 |
| **Urban100** | PI | 6.9435 | **3.6136** | 3.6980 | 3.7704 | 3.6523 | **3.5511** |
| | LPIPS | 0.4726 | 0.1433 | 0.1551 | **0.1229** | 0.1500 | **0.1184** |
| | PSNR | 23.145 | 24.013 | 24.397 | 24.360 | **25.464** | **24.799** |
| | SSIM | 0.9011 | 0.9364 | 0.9381 | 0.9453 | **0.9505** | **0.9481** |

optimization. We set the learning rates to $1 \times 10^{-4}$ for both generator and discriminator, and reduce them to half at 50k, 100k, 200k, 300k iterations. As for the trade-off parameters of losses, we follow the settings in [42] and set $\beta^I_{SR}$ and $\gamma^I_{SR}$ to 0.01 and 0.005, accordingly. Then we set the weights of gradient loss equal to those of image-space loss. Hence $\beta^{GM}_{SR} = 0.01$ and $\gamma^{GM}_{SR} = 0.005$. In terms of $\beta^{GM}_{GB}$, we set it to 0.5 for better performance of gradient translation. All the experiments are implemented by PyTorch [31] on NVIDIA GTX 1080Ti GPUs.

### 4.2. Results and Analysis

**Quantitative Comparison**: We compare our method quantitatively with state-of-the-art perceptual-driven SR methods including SFTGAN [41], SRGAN [27], ESRGAN [42] and NatSR [37]. Results of PI, LPIPS, PSNR and SSIM values are presented in Table 1. In each row, the best result is highlighted in red while the second best is in blue. We can see in all the testing datasets SPSR achieves the best PI and LPIPS performance. Meanwhile, we get the second best PSNR and SSIM values in most datasets. It is noteworthy that while NatSR gets the highest PSNR and SSIM values in all the datasets, our method surpasses NatSR by a large margin in terms of PI and LPIPS. Moreover, NatSR cannot achieve the second best PI and LPIPS values in any testing set. Thus NatSR is more like a PSNR-oriented SR method, which tends to produce relatively blurry results with high PSNR compared to other perceptual-driven methods. Besides, we get better performance than ESRGAN with only a little increment on network parameters in the SR branch. Therefore, the results demonstrate the superior ability of our SPSR method to obtain excellent perceptual quality and minor distortions simultaneously.

**Qualitative Comparison**: We also conduct visual comparison to perceptual-driven SR methods. From Figure 4 we see that our results are more natural and realistic than other methods. For the first image, SPSR infers sharp edges of the bricks properly, indicating that our method is capable of capturing structural characteristics of objects in images. In other rows, our method also recovers better textures than the compared SR methods. The structures in our results are clear without severe distortions, while other methods fail to show satisfactory appearance for the objects. Gradient maps for the last row are shown in Figure 5. We can see the gradient maps of other methods tend to have small values or contain structure degradation while ours are bold and natural. The qualitative comparison proves that our proposed SPSR method can learn more structure information from the gradient space, which helps generate photo-realistic SR images by preserving geometric structures.

**User Study**: We further perform a user study to evaluate visual quality of different SR methods. Detailed settings and results are presented in the supplementary material.

**Ablation Study**: We conduct more experiments on different models to validate the necessity of each part in our proposed framework. Since we apply the architecture of

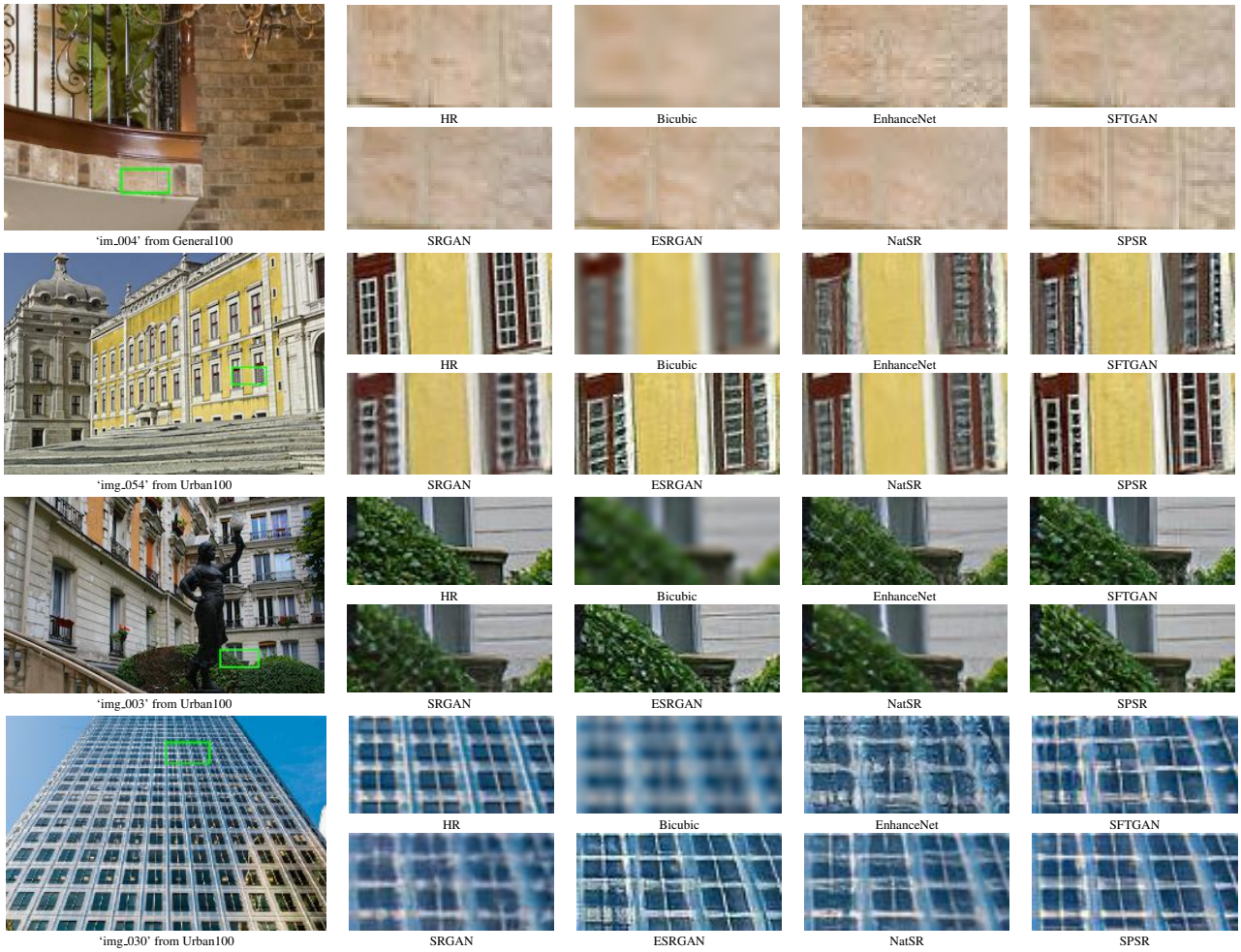



Figure 4. Visual comparison with state-of-the-art perceptual-driven SR methods. The results show that our proposed SPSR method significantly outperforms other methods in structure restoration while generating perceptual-pleasant SR images. Best viewed on screen.

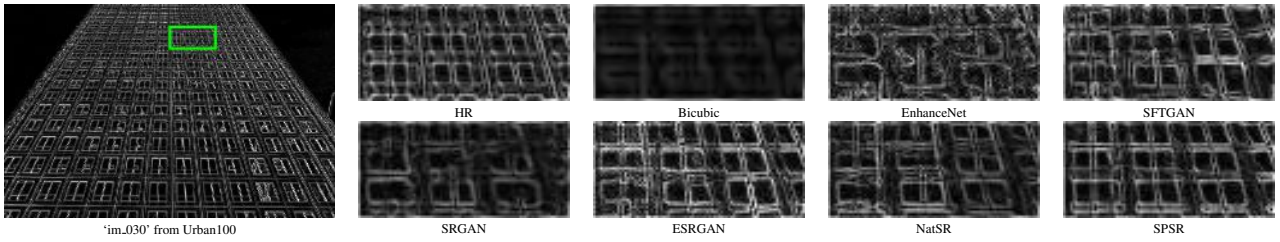


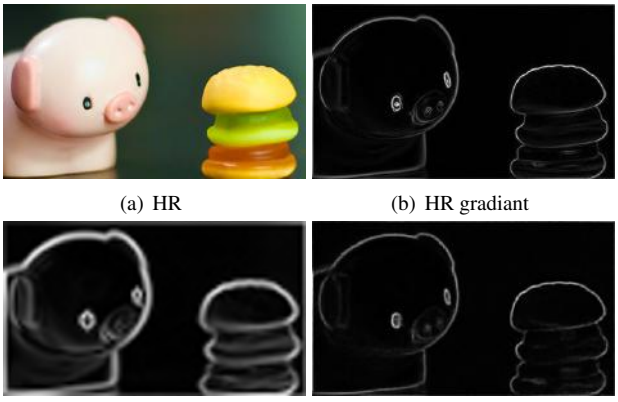

Figure 5. Comparison of gradient maps with state-of-the-art perceptual-driven SR methods. The proposed SPSR method can better preserve gradients and structures. Best viewed on screen.

ESRGAN [42] in our SR branch, we use ESRGAN as the baseline. We compare three models with it. The first one has the same architecture as ESRGAN without the gradient branch (GB) and is trained by both the image-space and gradient-space loss. The second one is trained without the gradient loss (GL), but has the gradient branch in the network. The third is our proposed SPSR model, utilizing both the gradient loss and the gradient branch. Quantitative comparison is presented in Table 2. It is observed that SPSR w/o GB has a significant enhancement on PI performance over ESRGAN, which demonstrates the effectiveness of the proposed gradient loss in improving perceptual quality. Besides, the results of SPSR w/o GL also show that the gradient branch can significantly help improve PI or PSNR while relatively preserving the other one. In terms of the complete model, we can see SPSR surpasses ESRGAN on all the measurements in all the testing sets. Therefore, the effectiveness of our method is verified clearly.

**Effects of the Gradient Branch**: In order to validate the effectiveness of the gradient branch, we also visualize the

Table 2. Comparison of models with different components. The best results are **highlighted**. SPSR w/o GB has better PI performance than ESRGAN in all the benchmark datasets. SPSR surpasses ESRGAN on all the measurements in all the testing sets.

| Method | Set14 | | | BSD100 | | | Urban100 | | |
|---|---|---|---|---|---|---|---|---|---|
| | PI | PSNR | SSIM | PI | PSNR | SSIM | PI | PSNR | SSIM |
| **ESRGAN** [42] | 2.926 | 26.276 | 0.778 | 2.479 | 25.317 | 0.651 | 3.770 | 24.360 | 0.945 |
| **SPSR w/o GB** | **2.864** | 26.027 | 0.785 | 2.370 | 25.376 | **0.659** | 3.604 | 23.939 | 0.940 |
| **SPSR w/o GL** | 3.028 | 26.547 | **0.794** | 2.456 | 25.214 | 0.647 | 3.605 | 24.309 | 0.942 |
| **SPSR** | 2.904 | **26.640** | 0.793 | **2.351** | **25.505** | 0.658 | **3.551** | **24.799** | **0.948** |



(a) HR (b) HR gradiant

(c) LR gradiant (Bicubic) (d) Output of the gradiant branch

Figure 6. Visualization of gradient maps ('im_073' from General100). The HR gradient map has thin outlines while those in the LR gradient map are thick. Our gradient branch is able to recover HR gradient maps with pleasant structures.

(a) Only the SR branch (b) Complete model

Figure 7. SR comparison of the models without and with the gradient branch ('baboon' from Set14). Images recovered by the complete model have clearer textures than those generated only by the features from the SR branch.

output gradient maps as shown in Figure 6. Given HR images with sharp edges, the extracted HR gradient maps may have thin and clear outlines for objects in the images. However, the gradient maps extracted from the LR counterparts commonly have thick lines after the bicubic upsampling. Our gradient branch takes LR gradient maps as inputs and produce HR gradient maps so as to provide explicit structural information as a guidance for the SR branch. By treating gradient generation as an image translation problem, we can exploit the strong generative ability of the deep model. From the output gradient map in Figure 6 (d), we can see our gradient branch successfully recover thin and structure-pleasing gradient maps.

We conduct another experiment to evaluate the effectiveness of the gradient branch. With a complete SPSR model, we remove the features from the gradient branch by setting them to 0 and only use the SR branch for inference. The visualization results are shown in Figure 7. From the patches, we can see the furs and whiskers super-resolved by only the SR branch are more blurry than those recovered by the complete model. The change of detailed textures reveals that the gradient branch can help produce sharp edges for better perceptual fidelity.

## 5. Conclusion

In this paper, we have proposed a structure-preserving super resolution method (SPSR) with gradient guidance to alleviate the issue of geometric distortions commonly existing in the SR results of perceptual-driven methods. We have preserved geometric structures in two aspects. Firstly, we build a gradient branch which aims to recover high-resolution gradient maps from the LR ones and provides gradient information to the SR branch as an explicit structural guidance. Secondly, we propose a new gradient loss to impose second-order restrictions on the recovered images. Geometric relationship can be better captured with both the image-space and gradient-space supervision. Quantitative and qualitative experimental results on five popular benchmark testing sets have shown the effectiveness of our proposed method.

## Acknowledgement

This work was supported in part by the National Key Research and Development Program of China under Grant 2017YFA0700802, in part by the National Natural Science Foundation of China under Grant 61822603, Grant U1813218, Grant U1713214, and Grant 61672306, in part by the Shenzhen Fundamental Research Fund (Subject Arrangement) under Grant JCYJ20170412170602564, and in part by Tsinghua University Initiative Scientfic Research Program.

## References


[1] Eirikur Agustsson and Radu Timofte. Ntire 2017 challenge on single image super-resolution: Dataset and study. In *CVPR*, pages 126–135, 2017.

[2] Asha Anoosheh, Torsten Sattler, Radu Timofte, Marc Pollefeys, and Luc Van Gool. Night-to-day image translation for retrieval-based localization. In *ICRA*, pages 5958–5964. IEEE, 2019.

[3] Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein gan. *arXiv preprint arXiv:1701.07875*, 2017.

[4] David Berthelot, Thomas Schumm, and Luke Metz. Began: Boundary equilibrium generative adversarial networks. *arXiv preprint arXiv:1703.10717*, 2017.

[5] Marco Bevilacqua, Aline Roumy, Christine Guillemot, and Marie-Line Alberi-Morel. Low-complexity single-image super-resolution based on nonnegative neighbor embedding. In *BMVC*, 2012.

[6] Yochai Blau, Roey Mechrez, Radu Timofte, Tomer Michaeli, and Lihi Zelnik-Manor. The 2018 pirm challenge on perceptual image super-resolution. In *ECCV*, pages 334–355. Springer, 2018.

[7] Hong Chang, Dit-Yan Yeung, and Yimin Xiong. Super-resolution through neighbor embedding. In *CVPR*, pages 275–282, 2004.

[8] Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang. Learning a deep convolutional network for image super-resolution. In *ECCV*, pages 184–199. Springer, 2014.

[9] Chao Dong, Chen Change Loy, and Xiaoou Tang. Accelerating the super-resolution convolutional neural network. In *ECCV*, pages 391–407. Springer, 2016.

[10] Claude E Duchon. Lanczos filtering in one and two dimensions. *Journal of applied meteorology*, 18(8):1016–1022, 1979.

[11] Raanan Fattal. Image upsampling via imposed edge statistics. *TOG*, 26(3):95, 2007.

[12] Gilad Freedman and Raanan Fattal. Image and video upscaling from local self-examples. *TOG*, 30(2):12, 2011.

[13] William T. Freeman, Thouis R. Jones, and Egon C. Pasztor. Example-based super-resolution. *CG&A*, 22(2):56–65, 2002.

[14] Daniel Glasner, Shai Bagon, and Michal Irani. Super-resolution from a single image. In *ICCV*, pages 349–356. IEEE, 2009.

[15] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *NIPS*, pages 2672–2680, 2014.

[16] Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron C Courville. Improved training of wasserstein gans. In *NIPS*, pages 5767–5777, 2017.

[17] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016.

[18] Jia-Bin Huang, Abhishek Singh, and Narendra Ahuja. Single image super-resolution from transformed self-exemplars. In *CVPR*, pages 5197–5206, 2015.

[19] Michal Irani and Shmuel Peleg. Improving resolution by image registration. *CVGIP*, 53(3):231–239, 1991.

[20] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *ECCV*, pages 694–711. Springer, 2016.

[21] Alexia Jolicoeur-Martineau. The relativistic discriminator: a key element missing from standard gan, 2018.

[22] Robert Keys. Cubic convolution interpolation for digital image processing. ieee trans acoust speech signal process. *TASSP*, 29:1153 – 1160, 01 1982.

[23] Jiwon Kim, Jung Kwon Lee, and Kyoung Mu Lee. Accurate image super-resolution using very deep convolutional networks. In *CVPR*, pages 1646–1654, 2016.

[24] Jiwon Kim, Jung Kwon Lee, and Kyoung Mu Lee. Deeply-recursive convolutional network for image super-resolution. In *CVPR*, pages 1637–1645, 2016.

[25] Kwang In Kim and Younghee Kwon. Single-image super-resolution using sparse regression and natural image prior. *TPAMI*, 32(6):1127–1133, 2010.

[26] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[27] Christian Ledig, Lucas Theis, Ferenc Huszár, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, et al. Photo-realistic single image super-resolution using a generative adversarial network. In *CVPR*, pages 4681–4690, 2017.

[28] Zhen Li, Jinglei Yang, Zheng Liu, Xiaomin Yang, Gwanggil Jeon, and Wei Wu. Feedback network for image super-resolution. In *CVPR*, pages 3867–3876, 2019.

[29] Fujun Luan, Sylvain Paris, Eli Shechtman, and Kavita Bala. Deep photo style transfer. In *CVPR*, pages 4990–4998, 2017.

[30] David R. Martin, Charless C. Fowlkes, Doron Tal, and Jitendra Malik. A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics. In *ICCV*, pages 416–425, 2001.

[31] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. In *NIPS-W*, 2017.

[32] Mohammad Saeed Rad, Behzad Bozorgtabar, Urs-Viktor Marti, Max Basler, Hazim Kemal Ekenel, and Jean-Philippe Thiran. Srobb: Targeted perceptual loss for single image super-resolution. *arXiv preprint arXiv:1908.07222*, 2019.

[33] Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*, 2015.

[34] Mehdi SM Sajjadi, Bernhard Scholkopf, and Michael Hirsch. Enhancenet: Single image super-resolution through automated texture synthesis. In *ICCV*, pages 4491–4500, 2017.

[35] Wenzhe Shi, Jose Caballero, Ferenc Huszár, Johannes Totz, Andrew P Aitken, Rob Bishop, Daniel Rueckert, and Zehan Wang. Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network. In *CVPR*, pages 1874–1883, 2016.

[36] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[37] Jae Woong Soh, Gu Yong Park, Junho Jo, and Nam Ik Cho. Natural and realistic single image super-resolution with explicit natural manifold discrimination. In *CVPR*, pages 8122–8131, 2019.
[38] Jian Sun, Zongben Xu, and Heung-Yeung Shum. Image super-resolution using gradient profile prior. In *CVPR*, pages 1–8. IEEE, 2008.
[39] Jian Sun, Zongben Xu, and Heung-Yeung Shum. Gradient profile prior and its applications in image super-resolution and enhancement. *TIP*, 20(6):1529–1542, 2010.
[40] Wei Wang, Ruiming Guo, Yapeng Tian, and Wenming Yang. Cfsnet: Toward a controllable feature space for image restoration. *arXiv preprint arXiv:1904.00634*, 2019.
[41] Xintao Wang, Ke Yu, Chao Dong, and Chen Change Loy. Recovering realistic texture in image super-resolution by deep spatial feature transform. In *CVPR*, pages 606–615, 2018.
[42] Xintao Wang, Ke Yu, Shixiang Wu, Jinjin Gu, Yihao Liu, Chao Dong, Yu Qiao, and Chen Change Loy. Esrgan: Enhanced super-resolution generative adversarial networks. In *ECCV*, pages 63–79. Springer, 2018.
[43] Zhou Wang, Alan C Bovik, Hamid R Sheikh, Eero P Simoncelli, et al. Image quality assessment: from error visibility to structural similarity. *TIP*, 13(4):600–612, 2004.
[44] Zhiwei Xiong, Xiaoyan Sun, and Feng Wu. Robust web image/video super-resolution. *TIP*, 19(8):2017–2028, 2010.
[45] Qing Yan, Yi Xu, Xiaokang Yang, and Truong Q Nguyen. Single image superresolution based on gradient profile sharpness. *TIP*, 24(10):3187–3202, 2015.
[46] Jianchao Yang, John Wright, Thomas Huang, and Yi Ma. Image super-resolution as sparse representation of raw image patches. In *CVPR*, pages 1–8, 2008.
[47] Jianchao Yang, John Wright, Thomas S Huang, and Yi Ma. Image super-resolution via sparse representation. *TIP*, 19(11):2861–2873, 2010.
[48] Wenhan Yang, Jiashi Feng, Jianchao Yang, Fang Zhao, Jiaying Liu, Zongming Guo, and Shuicheng Yan. Deep edge guided recurrent residual learning for image super-resolution. *TIP*, 26(12):5895–5907, 2017.
[49] Roman Zeyde, Michael Elad, and Matan Protter. On single image scale-up using sparse-representations. In *ICCS*, pages 711–730. Springer, 2010.
[50] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, pages 586–595, 2018.
[51] Yulun Zhang, Kunpeng Li, Kai Li, Lichen Wang, Bineng Zhong, and Yun Fu. Image super-resolution using very deep residual channel attention networks. In *ECCV*, pages 286–301, 2018.
[52] Yulun Zhang, Yapeng Tian, Yu Kong, Bineng Zhong, and Yun Fu. Residual dense network for image super-resolution. In *CVPR*, pages 2472–2481, 2018.
[53] Yu Zhu, Yanning Zhang, Boyan Bonev, and Alan L Yuille. Modeling deformable gradient compositions for single-image super-resolution. In *CVPR*, 2015.

# Supplementary Material

## A. User Study

We conduct a user study as a subjective assessment to evaluate the visual performance of different SR methods on benchmark datasets. HR images are displayed as references while SR results of our SPSR method, ESRGAN [42], NatSR [37] and SRGAN [27] are presented in a randomized sequence. Human raters are asked to rank the four SR versions according to the perceptual quality. Finally, we collect 1290 votes from 43 human raters. The summarized results are presented in Figure 8. As shown, our SPSR method gets much more votes of rank-1 than ESRGAN, NatSR and SRGAN. Meanwhile, most SR results of ESRGAN are voted the second best among the four methods since there are more structural distortions in the recovered images of ESRGAN than ours. NatSR and SRGAN fail to obtain satisfactory results. We think the reason is that they sometimes generate relatively blurry textures and undesirable artifacts. The comparison with the state-of-the-art GAN-based SR methods verifies the superiority of our proposed method in generating high-fidelity SR results.

## B. More Qualitative Results

We display more SR performance comparison with state-of-the-art SR methods including EnhanceNet [34], SFT-GAN [41], SRGAN [27], ESRGAN [42] and NatSR [37], as shown in Figure 9, 10, 11, 12 and 13. The results show our SPSR method performs better than other SR methods in recovering structural-pleasant and photo-realistic images. We also visualize the outputs of the gradient branch, as shown in Figure 14. We can see the gradient branch succeeds in converting LR gradient maps to the HR ones.

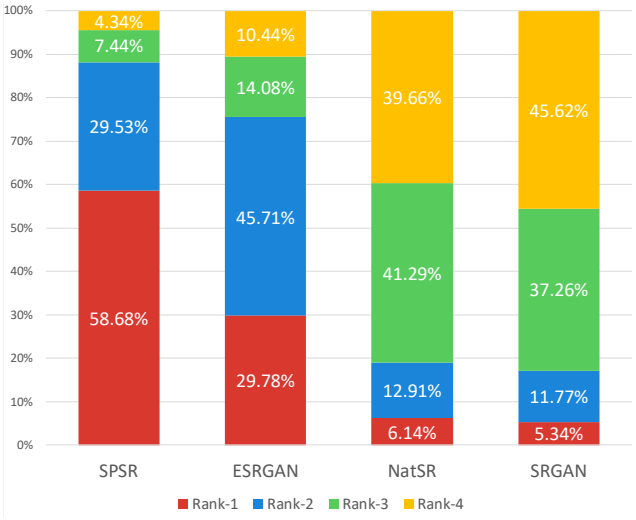



Figure 8. User study results of different GAN-based SR methods. Our SPSR method outperforms state-of-the-art SR methods in generating high-quality images.

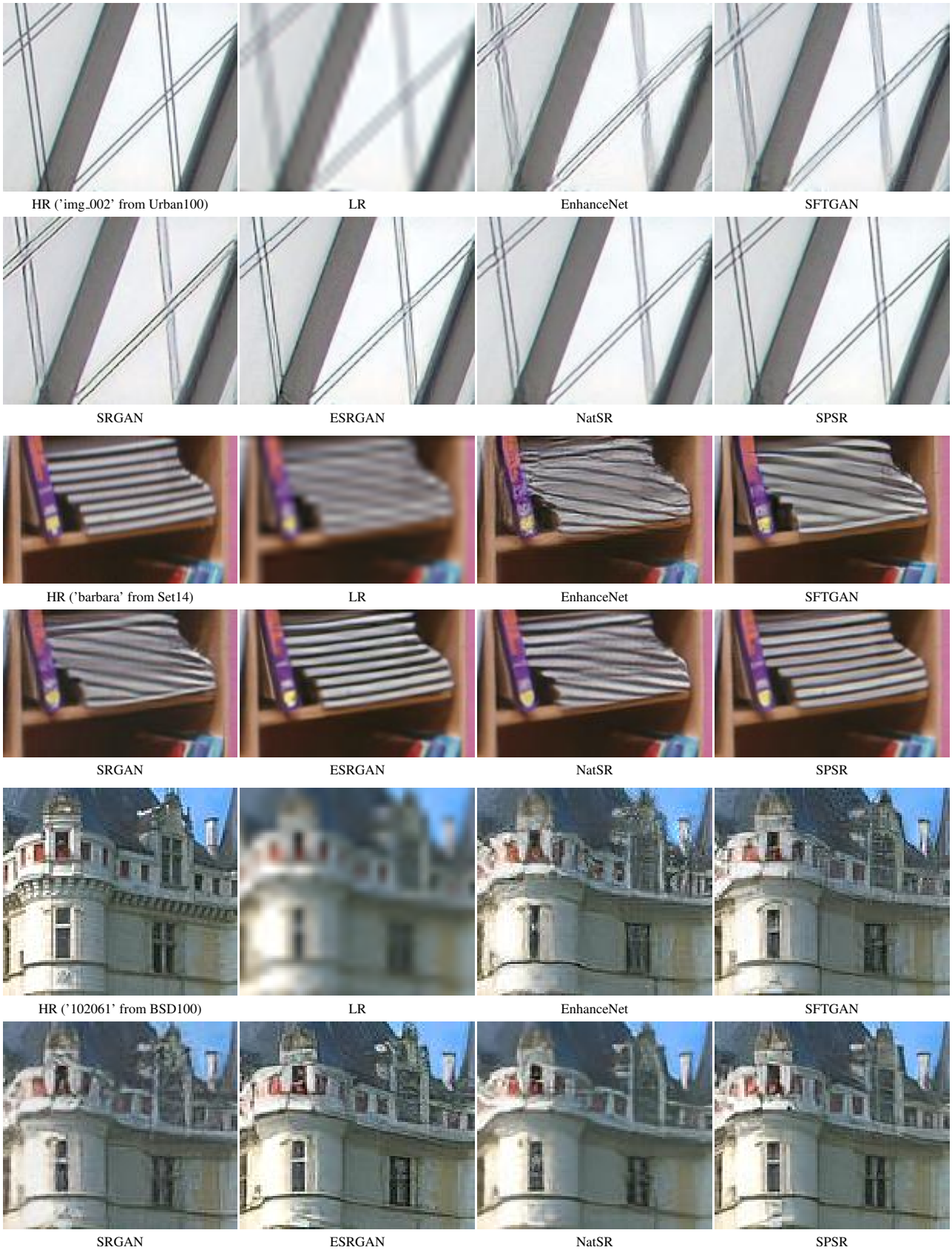



Figure 9. Visual comparison of SR performance with state-of-the-art SR methods.

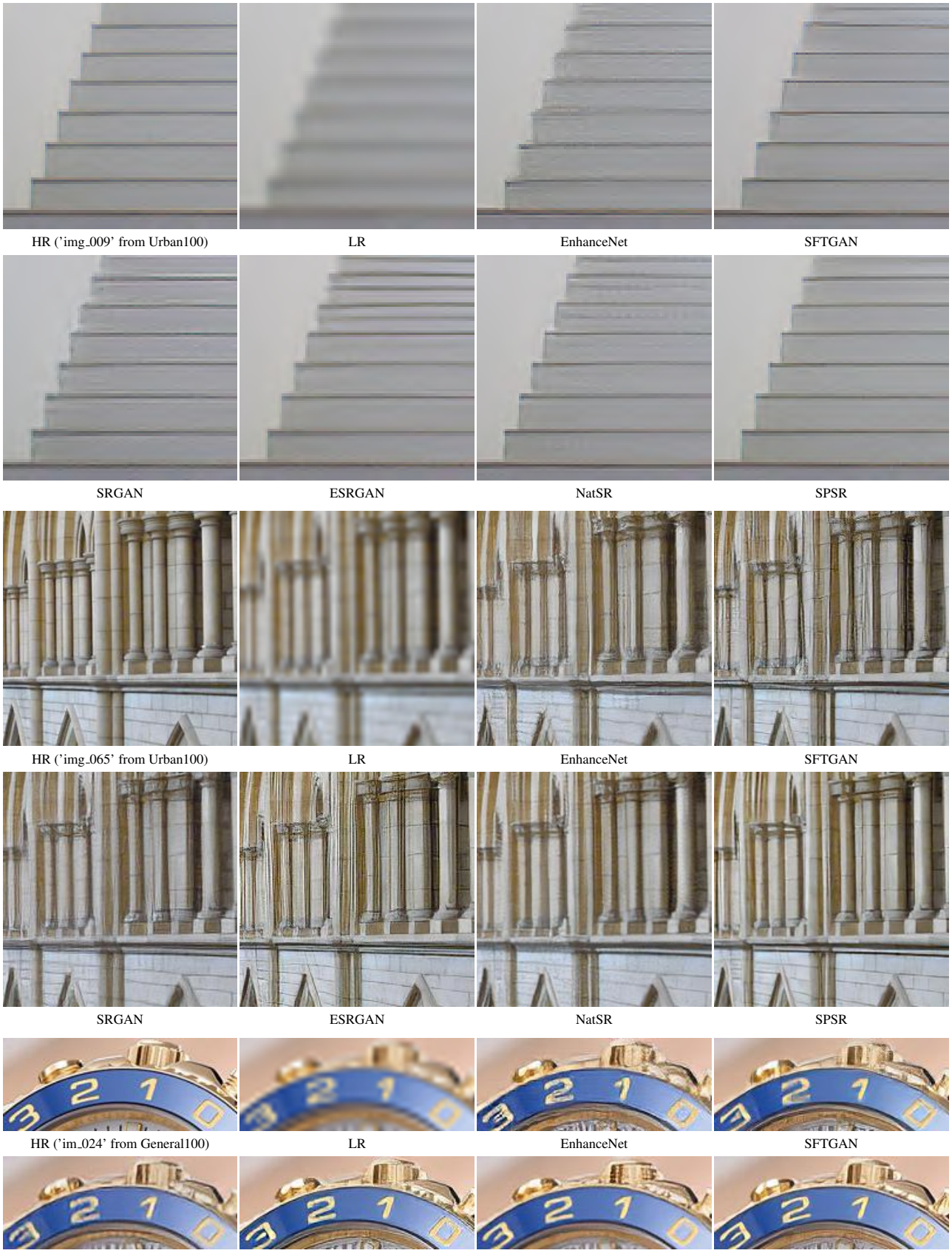



Figure 10. Visual comparison of SR performance with state-of-the-art SR methods.

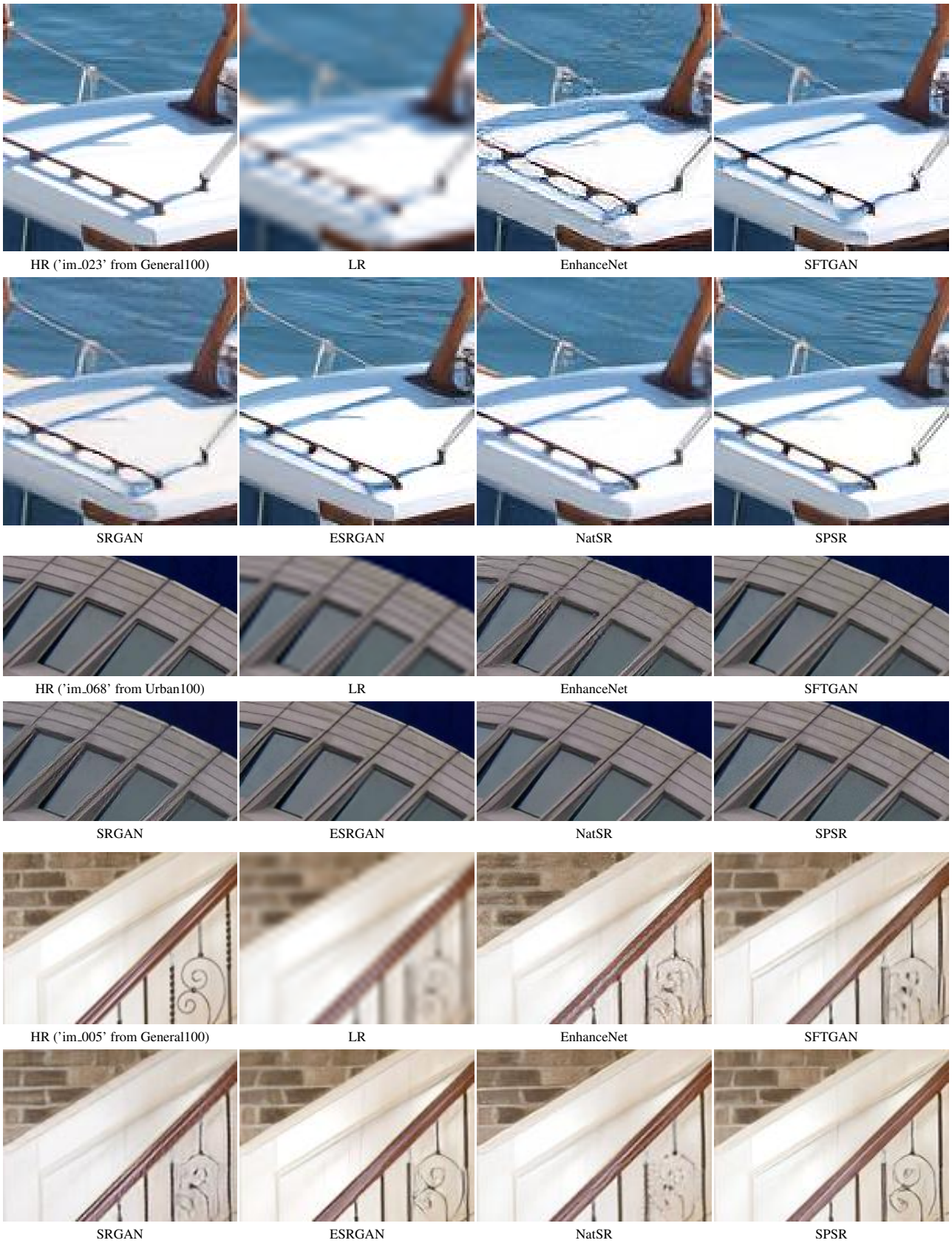
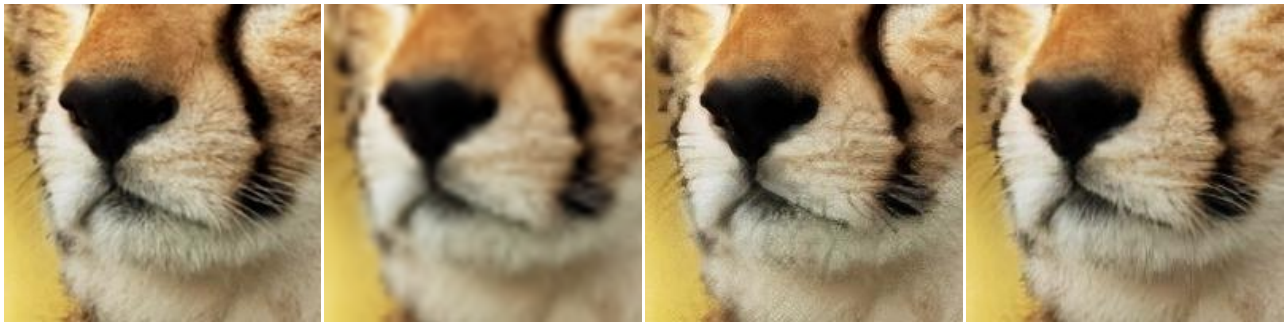



Figure 11. Visual comparison of SR performance with state-of-the-art SR methods.

HR ('img_077' from Urban100) LR EnhanceNet SFTGAN

SRGAN ESRGAN NatSR SPSR



HR ('im_008' from General100) LR EnhanceNet SFTGAN

SRGAN ESRGAN NatSR SPSR

Figure 12. Visual comparison of SR performance with state-of-the-art SR methods.

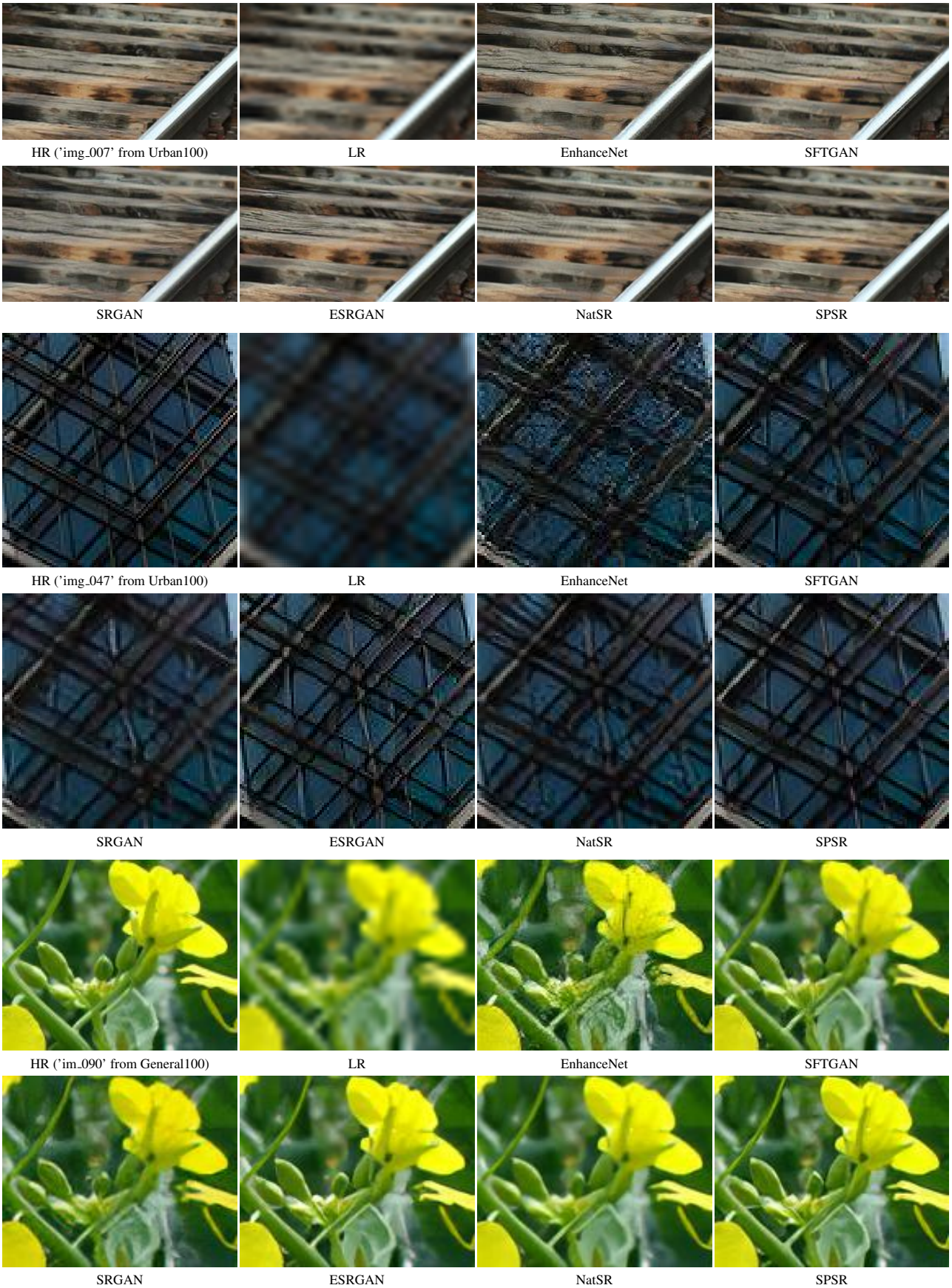
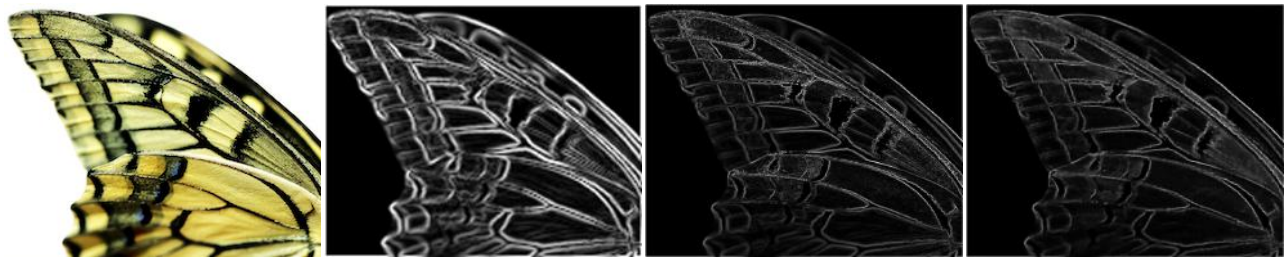
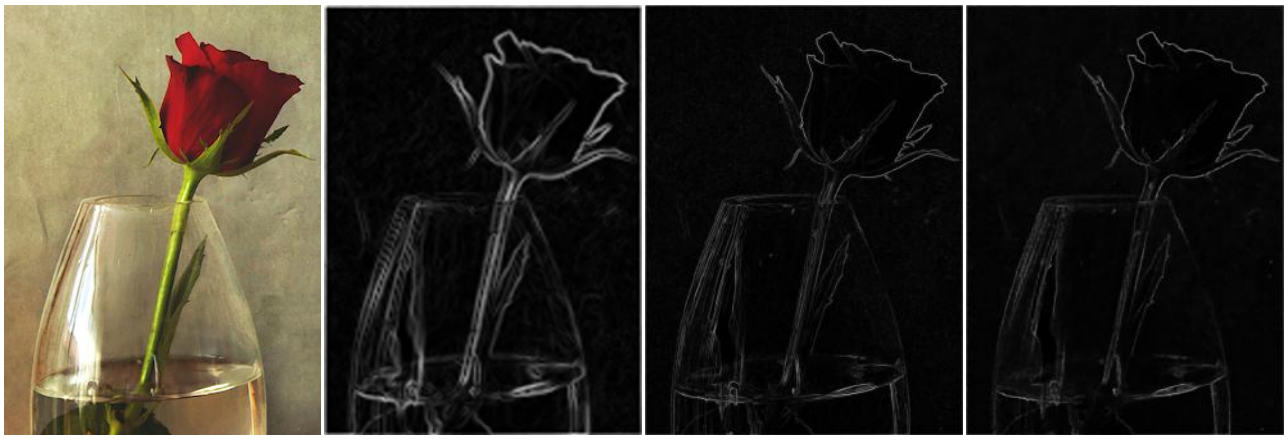


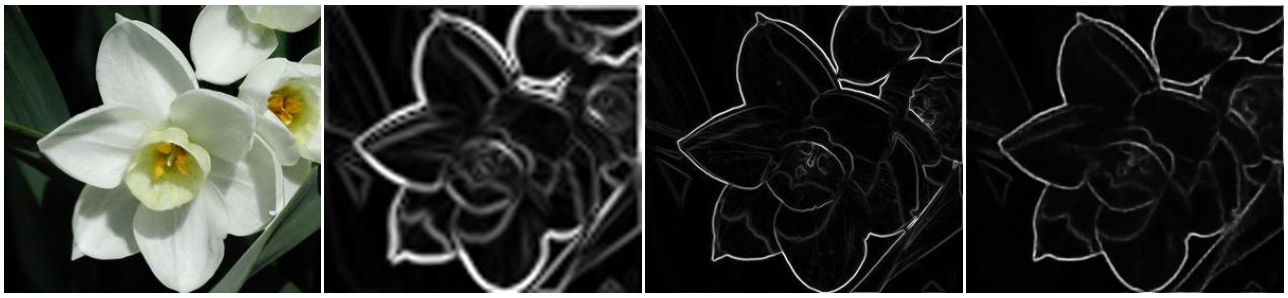

Figure 13. Visual comparison of SR performance with state-of-the-art SR methods.



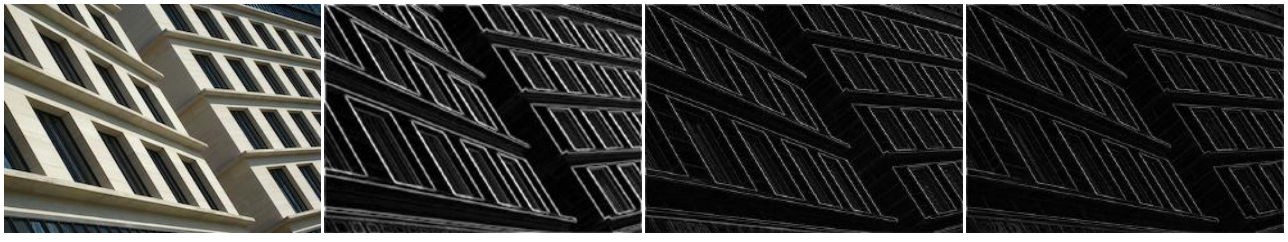
HR ('im_014' from General100) LR gradient (Bicubic) HR gradient Output of the gradient branch



HR ('im_026' from General100) LR gradient (Bicubic) HR gradient Output of the gradient branch



HR ('im_055' from General100) LR gradient (Bicubic) HR gradient Output of the gradient branch



HR ('img_025' from Urban100) LR gradient (Bicubic) HR gradient Output of the gradient branch

Figure 14. Visualization of gradient maps.