electronics

MDPI

Article

Unsupervised Single-Image Super-Resolution with Multi-Gram Loss

Yong Shi [1,2,3,4], Biao Li [1,2,3], Bo Wang [5,6], Zhiquan Qi [1,2,3,*] and Jiabin Liu [2,3]

[1] School of Economics and Management, University of Chinese Academy of Sciences, Beijing 101408, China
[2] Research Center on Fictitious Economy and Data Science, Chinese Academy of Sciences, Beijing 100190, China
[3] Key Laboratory of Big Data Mining and Knowledge Management, Chinese Academy of Sciences, Beijing 100190, China
[4] College of Information Science and Technology, University of Nebraska, Omaha, NE 68182, USA
[5] School of Information Technology and Management, University of International Business and Economics, Beijing 100029, China
[6] Department of Computer Science and Engineering, Texas A&M University, College Station, TX 77843, USA
* Correspondence: qizhiquan@ucas.ac.cn

Received: 20 June 2019; Accepted: 13 July 2019; Published: 26 July 2019

Abstract: Recently, supervised deep super-resolution (SR) networks have achieved great success in both accuracy and texture generation. However, most methods train in the dataset with a fixed kernel (such as bicubic) between high-resolution images and their low-resolution counterparts. In real-life applications, pictures are always disturbed with additional artifacts, e.g., non-ideal point-spread function in old film photos, and compression loss in cellphone photos. How to generate a satisfactory SR image from the specific prior single low-resolution (LR) image is still a challenging issue. In this paper, we propose a novel unsupervised method named unsupervised single-image SR with multi-gram loss (UMGSR) to overcome the dilemma. There are two significant contributions in this paper: (a) we design a new architecture for extracting more information from limited inputs by combining the local residual block and two-step global residual learning; (b) we introduce the multi-gram loss for SR task to effectively generate better image details. Experimental comparison shows that our unsupervised method in normal conditions can attain better visual results than other supervised SR methods.

Keywords: unsupervised single-image super-resolution; two-step super-resolution; multi-gram loss; global residual learning

1. Introduction

Super-resolution (SR) based on deep learning (DL) has received much attention from the community [1–7]. Recently, Convolutional neural networks (CNN)-relevant models have consistently resulted in significant improvement in SR generation. For example, the first CNN-based SR method SRCNN [4] generated more accurate SR images compared with traditional methods. In general, many high-resolution (HR)–low-resolution (LR) image pairs are the building blocks for DL-SR methods in a supervised way. The SR training uses the HR image as the supervised information to guide the learning process. Nevertheless, in practice, we barely collect enough external information (HR images) for training under severe conditions [8–10], e.g., medical images, old photos, and disaster monitoring images. On the other hand, most DL-SR methods train on the dataset with fixed kernel between HR and LR images. In fact, this fixed kernel assumption creates a fairly unrealistic situation limited in certain circumstances. When a picture violates the fixed spread kernel of training data, its final performance decreases in a large margin. This phenomenon is also highlighted in ZSSR [11]. In addition, if there are some artifacts, e.g., kernel noise or compression loss, a pre-trained DL model with a fixed kernel relationship will generate rather noisy SR images. As a result, we claim that we can turn to synthesis of the SR image with a single input, and it may become a solution to the problematic situation mentioned above.

Theoretically, SR is an ill-posed inverse problem. Many different SR solutions are suitable for one LR input. Intuitively, the more internal information of the LR input involves in the generation process, the better result can be expected. The changing route of DL-SR shows that various carefully designing strategies are being introduced to improve the learning ability. However, as a typical supervised problem, supervised DL-SR models train on the limited HR-LR image pairs. Model is restricted by the training data. In contrast, our method is conducted on single-input SR, i.e., designing a SR model for one image-input condition. We define the special condition as the unsupervised SR task following [11]. A new structure is proposed in our model. Moreover, to learn the global feature [12–14], we introduce the style loss to the SR task, i.e., the gram loss in the style transfer. Some experimental results show that the well-designed integrated loss can contribute to a better performance in the visual perception as depicted in [15].

Taking advantage of new structural design and loss functions, we can acquire considerably high-quality SR images both in the accuracy and the texture details. Specifically, the accuracy refers to the pixel alignment, which is commonly measured by the peak-signal-to-noise-ratio (PSNR) and the structural similarity index (SSIM) [2,4,5,7,16,17]. Moreover, the texture details are highlighted in some SR methods, such as [3,8,18,19], trying to generate satisfying images in visual perception by minimizing the feature distance between the SR image and its HR counterpart in some specific pre-trained CNN layers.

To sum up, in this paper, we propose a new unsupervised single-image DL-SR method with multi-gram loss (UMGSR) (Our code is available in the address: https://github.com/qizhiquan/UMSR). To address the aforementioned issues and improve visual performance, we dig three main modifications to the existing approaches. Firstly, we implement a specific unsupervised mechanism. Based on the self-similarity in [20], we denote the original input image as the G^{HR}. Then, the degradation operation is equipped to gain the corresponding G^{LR} counterpart. The training dataset is constituted with the G^{HR}–G^{LR} pairs. Secondly, we build a high-efficient framework with the residual neural network [21] as building blocks and introduce a two-step global residual learning to extract more information. The experimental results confirm that our approach performs well at the texture generation. Thirdly, we introduce the multi-gram loss following [22], which is commonly used in the texture synthesis. Accordingly, we form the loss function in UMGSR by combining the MSE loss, the VGG perceptual loss, and the multi-gram loss. Benefiting from these modifications, our model eventually achieves better performance in visual perception than both existing supervised and unsupervised SR methods. A comparison of SR images with different DL-SR methods is shown in Figure 1.

There are two main contributions in this paper:

- We design a new neural network architecture: UMGSR, which leverages the internal information of the LR image in the training stage. To stably train the network and convey more information about the input, the UMGSR combines the residual learning blocks with a two-step global residual learning.
- The multi-gram loss is introduced to the SR task, cooperating with the perceptual loss. In detail, we combine the multi-gram loss with the pixel-level MSE loss and the perceptual loss as the final loss function. Compared with other unsupervised methods, our design can obtain satisfying results in texture details and struggle for SR image generation similar to the supervised methods.

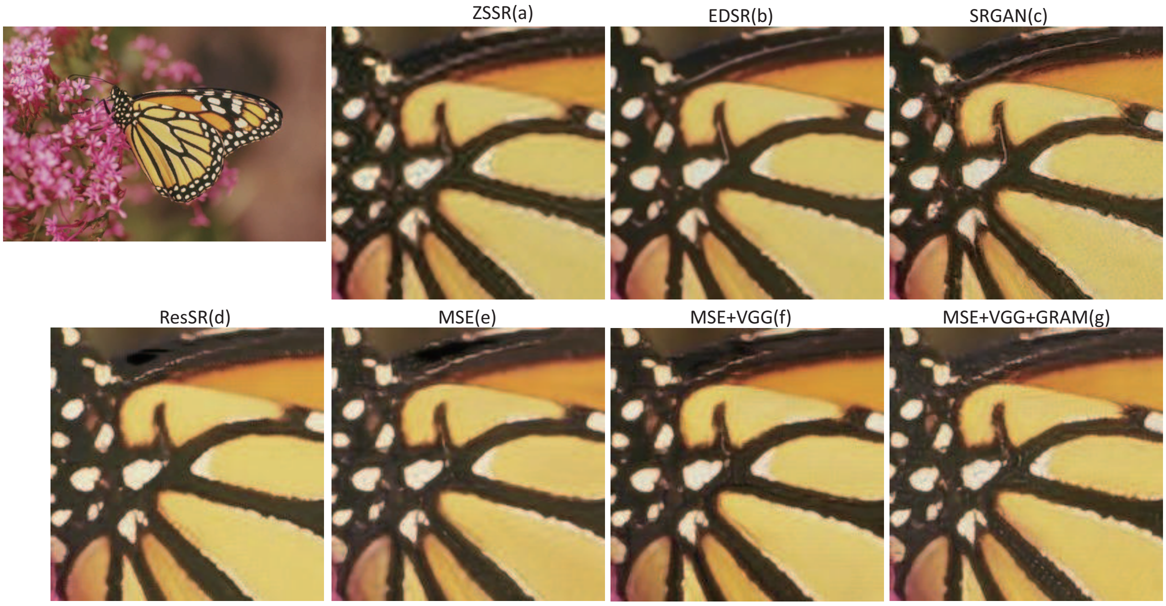


Figure 1. A comparison of some SR results. The figure shows the generation of ZSSR (an unsupervised DL-SR method), EDSR (a supervised method with best PSNR score), SRGAN (method good at the perceptual learning), ResSR (the generator of SRGAN), and our proposed method with three different loss functions. From the details, we can infer that more pleasant details are shown in the last pictures. The generations of different loss functions further provide change route of details.

2. Related Work

SR is one of basic computer vision tasks. In the realm of SR, there are mainly three distinct regimes: interpolation-based methods [23,24], reconstruction-based methods [25], and pairs-learning-based methods [1–5,7,11,20,26]. A lot of works are done to address this issue. like [27–29]. Recently, DL models achieve greatly success in many CV area, like [14,30–32]. In SR area, DL-SR methods become hugely successful, in terms of the performance both in accuracy and perceptual feeling. Most content achievements refer to outstanding DL-based approach and can be divided into three branches: supervised SR methods, unsupervised methods, and Generative Adversarial Networks (GAN) related methods.

Supervised SR methods. After AlexNet [33] firstly demonstrates the enormous advantage of DL over shallow methods in image classification, a large body of work applies deep CNN to traditional computer vision tasks. Regarding SR, the first DL-SR method is proposed by Dong et. at. in [4,34], which is a predefined upsampling method. It scales up of the LR image to the required size before training. Firstly, a traditional SR method (bicubic) is used to get the original scaled SR image. Then, a three layers CNN is employed to learn the non-linear mapping between the scaled SR image and the HR one. Noting that despite only three convolutional layers are involved, the result demonstrates a massive improvement in accuracy over traditional methods.

Later, researchers succeed in building sophisticated SR networks to strive for more accurate performance with relatively reasonable computation resource. For example, a new upsampling framework: the Efficient Sub-Pixel five layers Convolutional Neural Network (ESPCN), is proposed in [7]. Information of different layers is mixed to obtain the SR result. Meanwhile, the training process works with the small size LR input, and the scale-up layer is based on a simple but efficient sub-pixel convolution mechanism. Because most layers deal with small feature maps, the total computation complexity of ESPCN is considerably dropped. The sub-pixel scaling strategy is widely used in subsequent algorithms, such as SRGAN [3] and EDSR [1].

On the other hand, as mentioned in SRCNN, while it is a common sense that a deeper model accompanied with better performance, increasing the number of layers might result in non-convergence.

To bridge this gap, Kim et al. design a global residual mechanism following the residual neural network [21], to obtain a stable and deeper network. This mechanism eventually develops into two approaches: Very Deep Convolutional Networks (VDSR) [5] and Deeply Recursive Convolutional Network (DRCN) [35]. Due to the residual architecture, both networks can be stacked with more than 20 convolution layers, while the training process remains reasonably stable.

The following SR research mostly focuses on designing new local learning blocks. To building a deep and concise network, Deep Recursive Residual Network (DRRN) is proposed in [6], which replaces the residual block of DRCN with two residual units to extract more complex features. Similar to DRCN, by rationally sharing the parameters across different residual blocks, the total parameters of DRRN are controlled in a small number, while the network can be further extended to a deeper one with more residual blocks. In the DenseSR [36], new feature extracting blocks from DenseNet [37] contribute to fairly good results. To leverage the hierarchical information, Zhang et al. propose Residual Dense Block (RDB) in Residual Dense Network (RDN) [17]. Benefiting from the learning ability of local residual and dense connection, RDN achieves state-of-the-art performance. Besides, the Deep Back-Projection Networks (DBPN) [2] employs mutually up-down sampling stages and error feedback mechanism to generate more accurate SR image. Features of LR input are precisely learned by several repetitive up and down stages. DBPN attains stunning results, especially for large-scale factors, e.g., 8×.

Unsupervised SR methods. Instead of training on LR-HR image pairs, unsupervised SR methods leverage the internal information of single LR image. In general, there are a large body of classical SR methods follow this setting. For example, [38,39] make use of many LR images of the same scene but differing in sub-pixels. If the images are adequate, the point-spread function (PSF) can be estimated to generate the SR image. The SR generations are from a set of LR images with blurs, where pixels in the fixed patch following a given function. However, in [40], the maximum scale factor of these SR methods is proved to be less than 2. To overcome this limitation, a new approach trained with a single image is introduced in [20]. As mentioned in the paper, there are many similar patches of the same size or across different scales in one image. Then, these similar patches build the LR-HR image pairs, according to the single input and scaled derivatives for PSF learning. The data pre-processing in our work is similar to their idea. However, we adopt a DL model to learn the mapping between LR and SR images.

In addition, Shocher et al. introduce "Zero-Shot" SR (ZSSR) [11], which combines CNN and single-image scenario. Firstly, the model estimates the PSF as traditional methods. Then, a small CNN is trained to learn the non-linear mapping from the LR-HR pairs generated from the single-input image. In the paper, they prove that ZSSR surpasses other supervised methods in non-ideal conditions, such as old photos, noisy images, and biological data. Another unsupervised DL-SR model is the deep image prior [26], which focuses on the assumption that the structure of the network can be viewed as certain prior information. Based on this assumption, the initialization of the parameter serves as the specific prior information in network structure. In fact, this method suffers from over-fitting problem if the total epochs go beyond a limited small number. To our knowledge, the study of unsupervised DL-SR algorithm hardly receives enough attention, and there is still a big space for improvement.

GANs related methods. Generative Adversarial Networks (GANs) [41] commonly appears in image reconstruction tasks, such as [3,19,42,43], and is widely used for more realistic generation. The most important GAN-SR method is SRGAN [3], which intends to generate 4× upsampling photo-realistic images. SRGAN combines the content loss (MSE loss), perceptual loss [43], and adversarial loss in its last loss function. It can obtain photo-realistic images, although its performance on PSNR and SSIM indexes is relatively poor. In fact, our experiments also support their controversial discovery: a higher PSNR image does not have to deliver a better perceptual feeling. Besides, in [19], the FAN (face alignment) is introduced into a well-designed GAN model to yield better facial landmark SR images. Their experiments demonstrate significant improvements both in quantity and quality. For the restriction of facial image size, they use 16×16 as input to produce

64×64 output image. However, the FAN model is trained on a facial dataset, and it is only suitable for facial image SR problem. Inspired by the progress in GANs-based SR, we combine the SRGAN and Super-FAN in our architecture. We also make refined modification to address the unsupervised training issue.

3. Methodology

In this section, all details of the proposed UMGSR are shown in three folds: the dataset generation process, the proposed architecture, and the total loss. Referring to training DL-SR model upon unsupervised conditions, how to build the training data solely based on the LR image is the primary challenge to our work. Moreover, we propose a novel architecture to learn the map between generated $\hat{LR}$ and $\hat{HR}$ images. We also introduce a new multi-gram loss to obtain more spatial texture details.

3.1. The Generation of Training Dataset

How generating LR-HR image pairs from one LR input I^{in} is the fundamental task for our unsupervised SR model. Indeed, our work is a subsequent unsupervised SR learning following [11,20,44,45]. To generate satisfactory results, we randomly downscale I^{in} in a specific limited scale, which comes from the low visual entropy inside one image. Therefore, we obtain hundreds of different sizes I^{HR} and perform further operations based on these HR images.

Most supervised SR methods learn from dataset involving various image contents. The training data acts as the pool of small patches. There are some limitations for this setting: (1) the pixel-wise loss leads to over-smooth performance in the details; (2) supervised learning depends on specific image pairs and perform poorly when applied to significantly different images, such as old photos, noisy photos, and compressed phone photos; (3) no information of test image is involved in the training stage while it is crucial for the SR generation. Therefore, supervised SR models try to access the collection of external reference without the internal details of the test image. Figure 2 shows the mentioned drawbacks of supervised methods.

It can be inferred from the comparison that handrails of SR image in Glasner's [20] looks better than its counterpart of VDSR [5]. There are several similar repetitive handrails in the image, and details of different part or across various scale can be shared for their similarity. Training with these internal patches obtains better generations than the ones with external images. Normally, the visual entropy of one image is smaller than that of a set of different images [46]. Moreover, as mentioned in [11,46], lower visual entropy between images leads to better generation. Based on this consideration, learning with one image will result in an equal or better qualitative result than diverse $LR - HR$ image pairs. In our work, we continue this line of research by training with internal information, as well as incorporating more features. From Figure 1, we can see that our unsupervised method achieves a similar result as the state-of-the-art SR method in common conditions. For non-ideal images, it performs better.

Normally, the objective of SR task is to generate I^{SR} images from I^{LR} inputs, and information of I^{HR} acts as the supervised information during training. However, there are no or few available I^{HR} images for training in some specific conditions. Unsupervised learning seems to be a decent choice. In this circumstance, how to build the HR-LR image pairs upon a single image is a fundamental challenge. In our work, we formulate the dataset from the LR image by downsampling operation and data enhancement strategy. This maximized use of internal information contributes to a better quality of I^{SR}. Based on the generated training dataset, the loss function is shown as:

$$Loss = \arg\min \frac{1}{N} \sum_{i=1}^{n} (I_i^{HR} - G_i^{SR}(I^{LR})) \tag{1}$$

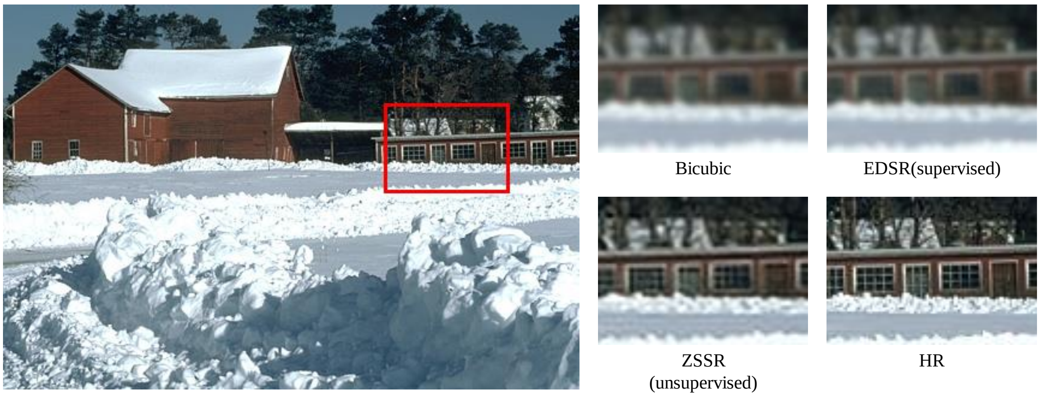


Figure 2. The comparison of supervised and unsupervised SR learning under "non-ideal" downscaling kernals condition. The unsupervised DL-SR method (ZSSR) firstly estimate the PSF, and learning internal information by a small CNN. The supervised method is one of the best ones named EDSR which is trained by a lot of image pairs. The comparing result shows that the unsupervised method surpasses the supervised method in the repetitive details, which potentially indicates the validity of internal recurrence for SR generation.

To obtain a comprehensive multi-scale dataset, we implement the data augmentation strategy on input image which is further down-scaled in a certain range. The process is in following. Firstly, an input image I acts as the I^{HR} image father. To use more spatial structure information, we introduce a down-scaled method to produce various different scaled HR images $I_i^{HR}, i = 1, 2, \cdots, n$, which are dealt with several different ratios. Secondly, we further downscale these I_i^{HR} with a fixed factor to get their corresponding LR images I_i^{LR} (i = 1,2,. . . ,n). Lastly, all these image pairs are augmented by rotation and mirror reflections in both vertical and horizontal directions. The final dataset contains image pairs with different shapes and contents. More information about the change of pixel alignment comes from a variety of scale images. In summary, all training pairs contain similar content architecture. Hence, the more pixel-level changing information among images of different sizes is involved, and then the better result will be yielded.

3.2. Unsupervised Multi-Gram SR Network

Based on ResSR, our model incorporates with a two-step global learning architecture inspired by [19]. Some specific changes are implemented for the specific of unsupervised SR purposes. Architectures of our UMGSR, ResSR, and Super-FAN are shown in Figure 3.

There is limited research on unsupervised DL-SR. To our knowledge, ZSSR [11] obtains a significant success in accurate pursuing route. They introduced a smaller and simpler CNN SR image-special model to obtain SR upon smaller diversity I_i^{HR} and I_i^{LR} from the same father image than any supervised training image pair. They announced that a simple CNN was sufficient to learn the SR map. At the same time, to some extent, the growth track of better PSNR supervised method indicates an obvious affinity between the network complexity and the SR generation accuracy. For example, EDSR [1] reports that their significant performance is improved by extending the model size. Therefore, we propose a more complex unsupervised model—UMGSR—shown in Figure 3c.

The total architecture of UMGSR. Generally speaking, the SR network can be divided into several blocks according to the diverse image scales during training. Taking 4× for example, there are three different inner sizes: the original input, the 2× up-scaling, and the 4× up-scaling. For simplicity, we define these intermediate blocks as L_{s1}, L_{s2}, and L_{s4}. Several blocks are stacked to learn the specific scale information in the corresponding stage. Then, ResSR leverages 16 residuals as L_{s1} for hierarchical convolution computation. The final part contains a 2× scaled block L_{s2} and a final 4× scaled one L_{s4}. In general, the total architecture of ResSR can be denoted as $16 - 1 - 1$ (i.e., $L_{s1} - L_{s2} - L_{s4}$).

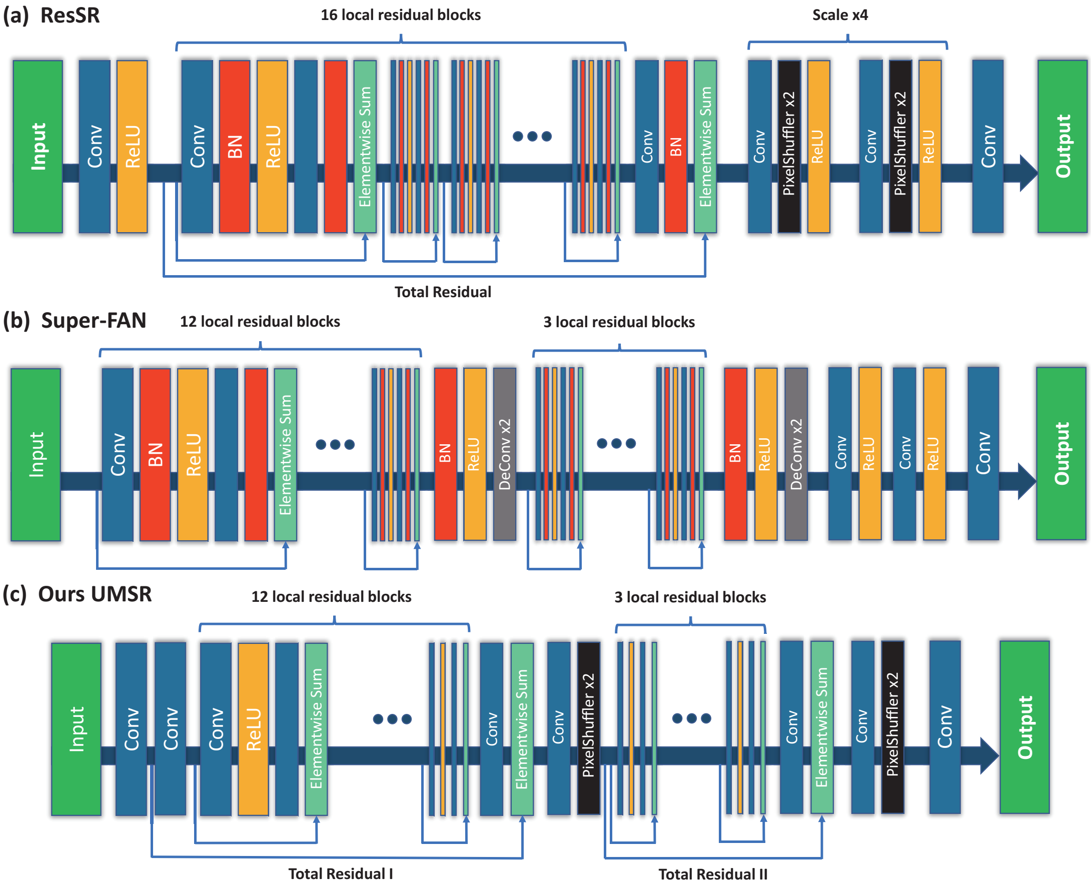


Figure 3. The architectures of ResSR, Super-FAN, and our UMGSR. (**a**) ResSR; (**b**) Super-FAN; (**c**) Ours UMSR.

From the comparison in Figure 3a–c, the architectures of three methods are: $16-1-1$, $12-4-2$, and $12-4-2$ respectively. The first part of the network contains one or two layers to extract features from the original RGB image. To this end, former methods mostly use one convolutional layer. By contrast, we use two convolutional layers for extracting more spatial information as in DBPN [2]. The first layer leverages a 3×3 kernel to generate input features for residual blocks. It is worth pointing out that there are more channels in the first layer for abundant features. For the purpose of acting as a resource of global residual, a convolutional layer with a 1×1 kernel is applied to resize layers same as the output of branch. For middle feature extracting part, the total residual blocks in all three models are similar. The main difference refers to the number of scaled feature layers. In fact, as pointed out in super-FAN, only using a single block at higher resolutions is insufficient for sharp details generation. Based on super-FAN, we build a similar residual architecture for a better generation. In detail, the middle process is separated into two sub-sections, and each subsection focuses on a specific $2\times$ scaled information learning. Inheriting the feature from the first part, layers in the first subsection extract features with the input size. Because more information of the input is involved here, more layers (12 layers) are employed in the first subsection, which aims at extracting more details of the image and producing sharper details. In contrast to the first subsection, the second one contains three residual blocks for further $2\times$ scale generation.

Global residual learning. Another important change is a step-by-step global residual learning structure. Inspired by ResNet, VDSR [5] firstly introduces global learning in SR, which succeeds in steady training a network with more than 20 CNN layers. Typically, the global learning can transmit the information from the input or low-level layer to a fixed high-level layer, which helps solve the problem of dis-convergence. Most of the subsequent DL-SR models introduce global learning strategy in their architectures to build a deep and complicated SR network. As shown in Figure 3a, the information from the very layer before the local residual learning and the last output layer of the local residual learning are combined in the global residual frame. However, only one scaling block for SR image generation is not enough for the large-scale issue. Therefore, in UMGSR, we arrange the global residual learning in each section: two functional residual blocks with two global residual learning frames. In fact, the first global learning fulfills stable training, and the closely adjacent second section can leverage similar information of the input image.

Local residual block architecture. Similar to SRGAN, all local parts are residual blocks which has proved to achieve better features learning results. During the training stage, we also explore the setting as in EDSR [1] abandoning all batch normalization layers. In general, the local residual block contains two 3 × 3 convolutional layers and a ReLU activation layer following each of them. Results of ResSR and EDSR elucidate the superior learning ability of this setting.

3.3. Pixel, Perceptual, and Gram Losses

In the realm of SR, most DL-SR methods train models with the pixel-wise MSE loss. Because there is direct relationship between MSE loss and standard PSNR index which commonly measures final performance. In [43], a novel perceptual loss is proposed to learn texture details. The new loss calculates Euclidean distance between two specially chosen layers from a pre-trained VGG19 [47] network. In SRGAN [3], the perceptual loss is firstly introduced to SR, and it shows great power in the generation of photo-realistic details.

Another loss for feature learning is the gram loss [13] which is widely used in the realm of style transfer. Gram loss performs as a global evaluating loss, which measures the style consistent. To extract more information about spatial structure, we use multi-gram loss in this paper. Ultimately, the loss function of UMGSR combines MSE loss, perceptual loss, and the multi-gram loss. More details are shown in the followings.

Pixel-level loss. Pixel-level loss is used to recover high-frequency information in I_i^{SR} with supervised I_i^{HR}. Normally, traditional l_1 or l_2 norm loss is widely used in DL-SR model, and they can produce results with satisfactory accuracy. In our UMGSR, the MSE loss is also introduced as the principle pixel-level loss for high accuracy. It is defined as:

$$Loss_{mse} = \frac{1}{s^2WH} \sum_{w=1}^{sW} \sum_{h=1}^{sH} (I_{wh}^{HR} - G_{P_G}(I_{wh}^{LR})^2, \tag{2}$$

where W and H are shape factors of input, and s is the scale factor.

The MSE loss contributes to finding the least distance in pixel-level among all possible solutions. When measuring the accuracy, models achieve the best PSNR and SSIM without using other loss. However, the I_s^{SR} suffers from the over-smooth issue, which leads to an unreal feeling in visual. A detailed illustration will be shown in the experimental part. To deal with this problem, we further propose perceptual loss and multi-gram loss.

Perceptual loss. To obtain more visual satisfying details, we apply the perceptual loss [43] as in SRGAN [3], which minimizes the Euclidean distance of a pre-trained VGG19 [47] layer between the corresponding HR and SR images. It aims at better visual feeling results, as well as reducing of PSNR. To facilitate the understanding, we illustrate the architecture of VGG19 in Figure 4.

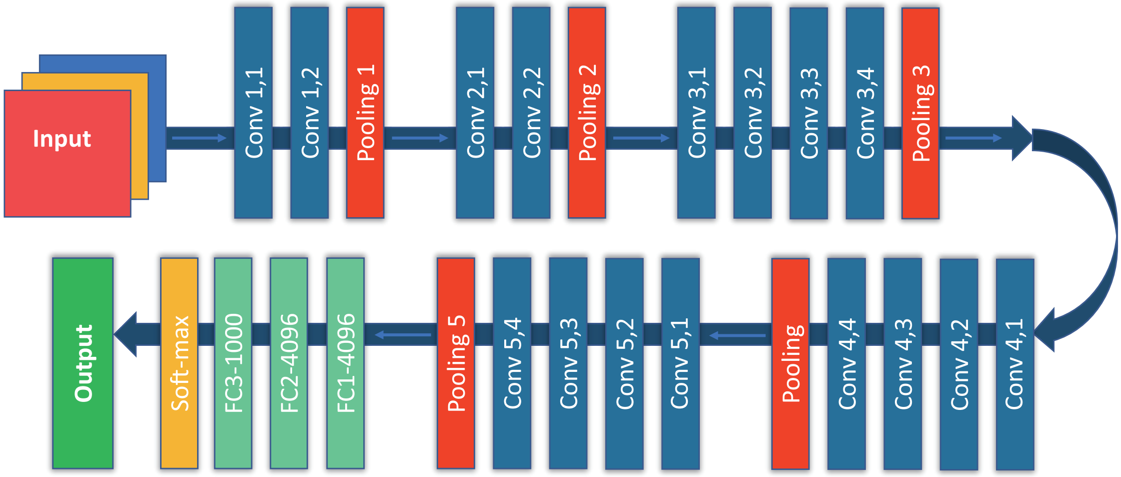


Figure 4. The architecture of VGG19.

In SRGAN, only one specified layer of VGG19 is involved in the perceptual loss, i.e., $VGG_{5,4}$ (the fourth convolution before the fifth pooling layer). Different layers of the network represent various levels of feature. In other words, the former part learns intensive features, and the latter one learns larger coverage information. As a result, we argue that one layer for perceptual loss is not enough. To fix this weakness, we propose a modified perceptual loss by mixing perceptual losses in several different layers of VGG19. In our experiments, we use the combination of $VGG_{2,2}$, $VGG_{2,3}$, $VGG_{3,4}$, and $VGG_{5,4}$ with different trade-off weights, i.e.,

$$\begin{cases} loss_p = \alpha_1 V_{2,2} + \alpha_2 V_{3,2} + \alpha_3 V_{4,3} + \alpha_4 V_{5,4}, \\ \sum_{i=1}^{4} \alpha_i = 1. \end{cases} \tag{3}$$

In fact, this new loss helps us abstract feature information in different feature sizes. Although it is proved in [7] that the perceptual loss in high-level layer promotes better texture details, we still insist that the training of DL-SR network is a multi-scale learning process, and more information involved can potentially lead to better results. During experiments, we propose a perceptual loss to generate visual transition details from high-frequency information.

Multi-gram loss. In style transfer, the gram matrix measures the relationship among all inner layers in a chosen channel. It supplies the global difference information of all image features. The gram loss is first introduced to DL in [13], to train a DL network with gram loss as a style loss and MSE loss as a content loss between two images. In SR, I_i^{HR} and I_i^{SR} share similar spatial architecture and features. More spatially invariant can be extracted by the feature correlations in different sizes. Compared with style transfer, we introduce the multi-gram loss [22] in UMGSR to generate better visual details as [22], which first proposes the multi-gram loss from the Gaussian pyramid in a specific layer. Our redesign of the multi-gram loss for the SR purpose is shown as follows:

$$\begin{cases} G_{ij}^{r,s} = \frac{1}{M_r N_r} \sum_f F_{if}^{r,s} F_{jf}^{r,s}, \\ E_r^s = \sum_{ij} \left(\hat{G}_{ij}^s - G_{ij}^s \right)^2, \\ L(\hat{\vec{x}}, \vec{x}) = \sum_{s=0}^{S-1} v_s \sum_{r=0}^{R-1} w_r E_r^s. \end{cases} \tag{4}$$

In detail, the first function calculates the gram matrix in a specific layer. All i, j, r, s represent different feature maps: i, j in the r^{th} layer and the s^{th} scale octave of the Gaussian pyramid. The second function measures the gram loss between the source image and its counterpart. The last function refers to the specially chosen layers, where we expect to extract the gram loss. The values of v and w are chosen from 1 or 0, to keep or abandon the gram loss of one certain scale layer, respectively.

The multi-gram loss determines the overall global texture in image compared to the perceptual loss on local features. Each of them can be served as the complementary role to another. The experiments show their positive effect on the details of the final SR output. In general, the final loss of UMGSR is constituted by summing up all the three losses with specific trade-off factors as:

$$Loss_x = \alpha Loss_{mse} + \beta Loss_p + \gamma Loss_{gram}. \tag{5}$$

4. Experiments

In this part, we conduct contrast and ablation experiments to evaluate our proposed UMGSR. All of our models are trained on a NVIDIA TITAN XP GPU with 4× scale factor. There are three parts as follows:

4.1. Setting Details

Because just one image acts as the input of UMGSR, we choose all input images I^{in} from three different benchmark datasets (Set14 dataset [48], DIV2K dataset [49], and PIRM dataset [15]), to conduct a fair comparison with other supervised and unsupervised methods. The images with content consistent to various complicated conditions are qualified as the realistic ones.

Training setting details. As mentioned in the methodology part, we firstly apply the data augment strategy to form the training dataset from I^{in}. To obtain I_i^{HR} (i=1,2,…,n), we randomly scale I^{in} in the range of 0.5 to 1, following with rotation on I_i^{HR} in both horizontal and vertical directions. In addition, we do not apply random cropping, so that more information of I^{in} can be kept. The initial learning rate is set to be 0.001, with half reducing when remaining epochs are half down. We perform Adam ($\beta_1 = 0.9$, $\beta_2 = 0.999$) to optimize the objective. The patch size is 30×30, and the corresponding HR size is changed to 120×120. The I_i^{LR} (i=1,2,…,n) images are with smaller size since they are $4\times - 8\times$ down-scaled from the I^{in} images. We set the total training epochs as 4000.

Ablation setting. In the following part, we demonstrate the influence of proposed changes in UMGSR by ablation analysis. To this end, firstly, we train our model only with MSE loss. Secondly, we use both the MSE loss and the perceptual loss. Here, we also consider the comparison between single perceptual loss and the incorporating one to evaluate its influence. Finally, we investigate the performance with the total loss, combining the MSE, the perceptual, and the multi-gram loss. Except for the loss function, all other settings are kept consistently. We parallelly compare the generations of UMGSR (with different loss functions and structures), EDSR(https://github.com/thstkdgus35/EDSR-PyTorch), SRGAN(https://github.com/tensorlayer/srgan), and ZSSR (https://github.com/assafshocher/ZSSR). All generations are obtained by the pre-trained models from the url links. All results are compared in PSNR (Y channel), which measures the accuracy in pixels, and another total distribution index: the spectral image. Moreover, we further present the detail comparison of the same patch from all generations.

Structure setting. UMGSR with 15 residual blocks is shown in Figure 3. In detail, the former 12 blocks are used to extract the first 2× features from the input. The remaining three residual blocks inherit information from previous 2× scaled blocks and achieve 4× up-scaling. All filter sizes equal to 30×30, and all residual blocks include 64 channels for feature learning in contrast to 256 channels in the deconvolutional part. We train the model with the 1008 HR-LR image pairs from one image.

4.2. Ablation Experiments

Training when β and γ are equal to zero. As most DL-SR methods, we use the MSE loss as the basic loss function. In this setting, our model is similar to the ResSR except for single difference in the total architecture. To show changes of new structure, we compare them with only structure difference. The final results of these two methods are shown in Figure 5.

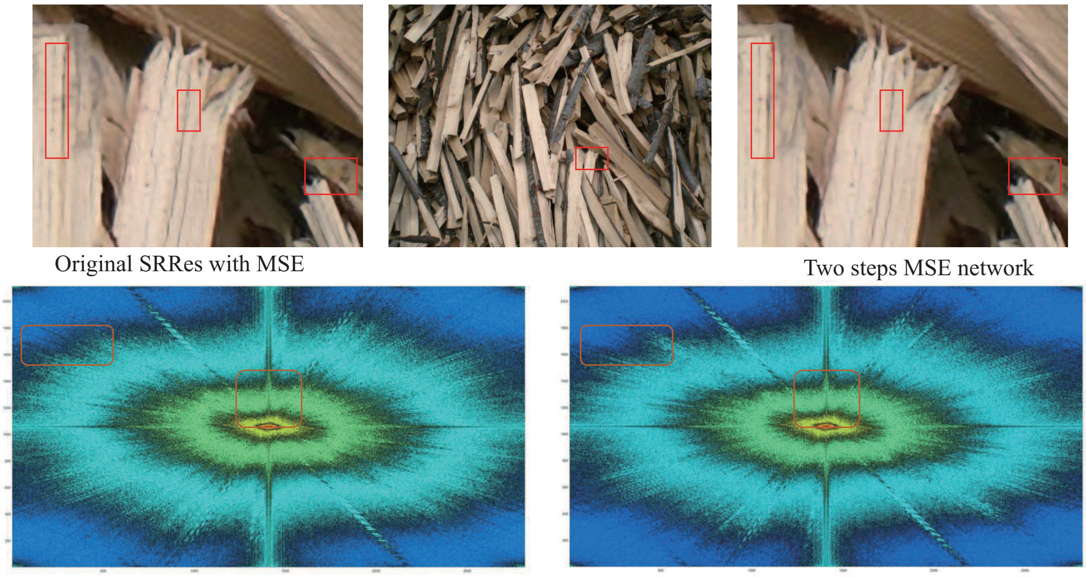
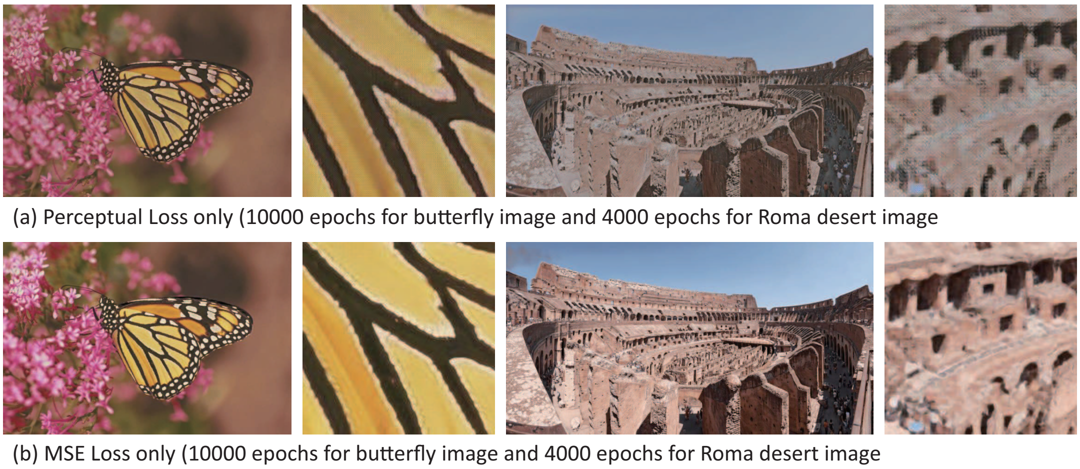


Figure 5. Details comparison between SRRes and two-step learning with MSE loss. From the left box, we can acquire that clear growth ring is generated with new structure. It is also shown in the spectral image.

From the results, we can see that our two-step network produces pictures with more natural feeling than ResSR. In addition, spectral comparison in Figure 5 shows that the two-step network generates more accurate features. There is less blur information in the red rectangular area where two-step strategy is used.

Training when γ equals to zero. In this part, we introduce the perceptual loss to the loss function. To be specified, layers $VGG_{2,2}$ and $VGG_{4,3}$ of VGG19 are used in the final loss function by fixing $\alpha_1 = 0.3$ and $\alpha_3 = 0.7$ in (3). Here, to comprehensively distinguish the effect of perceptual loss, we display the comparison between training with only the perceptual loss and with the combination of MSE and perceptual loss in Figure 6.


(a) Perceptual Loss only (10000 epochs for butterfly image and 4000 epochs for Roma desert image

(b) MSE Loss only (10000 epochs for butterfly image and 4000 epochs for Roma desert image

Figure 6. The comparison on perceptual loss and MSE loss: (**a**)just perceptual loss. (**b**)only MSE loss.

From the detail contrast, we can tell that with single perceptual loss, many features in local block are missing. In our opinion, this phenomenon is due to the upsampling stage where the input must be enlarged by Bicubic to the required input size of VGG network, i.e., 224 × 224. However, the I_{SR} and I_{HR} images in UMGSR is 120 × 120. As a result, a lot of unfitting information appears in up-scaled images. This local mismatching information further results in poor generations.

Training with all loss settings. In this part, we use the loss by incorporating the MSE loss, the perceptual loss, and new multi-gram loss. With the multi-gram loss, the network learns feature map in both global and local aspects. Because multi-gram loss measure spatial style losses, it leads to better visual feeling results both in details and shapes. Referring to super-parameters, $\alpha = 1$ and $\beta = \gamma = 2 \times 10^{-6}$. This setting is proved to be useful by SRGAN. In general, the final loss function is:

$$Loss_{total} = Loss_{mse} + 2 \times 10^{-6} Loss_{vgg_{5,4}} + \underbrace{2 \times 10^{-6} Loss_{gram}}_{\sum_{i=1}^{5}(Gram_{2,1}^{s_i} + Gram_{3,2}^{s_i})/5} . \quad (6)$$

The multi-gram loss is somehow similar to the perceptual loss. Both learn loss from inner layers of a pre-trained VGG network with the final SR image and its corresponding HR image as the inputs. For multi-gram loss, the $VGG_{2,1}$ and $VGG_{3,2}$ are chosen to be the specified loss layers. All chosen layers are down-scaled to five pyramid sizes for spatial adaption. The size of the chosen layers must be large enough. Then, five different sub-layers-like pyramid structure are used to calculate gram losses as mentioned in Section 3.3. Similar to the perceptual loss, extra noise appears in the SR results if the model trained only with multi-gram loss.

The final PSNR of images are summarized in Table 1, and the visual comparison is shown in Figure 7. With the introduction of multi-gram loss, more pleasant features appear in generations, which can be clearly observed in Figure 1. Furthermore, the MSE changing chart shows the advantage of final loss (combination of MSE, perceptual loss, and multi-gram loss) in Figure 8.

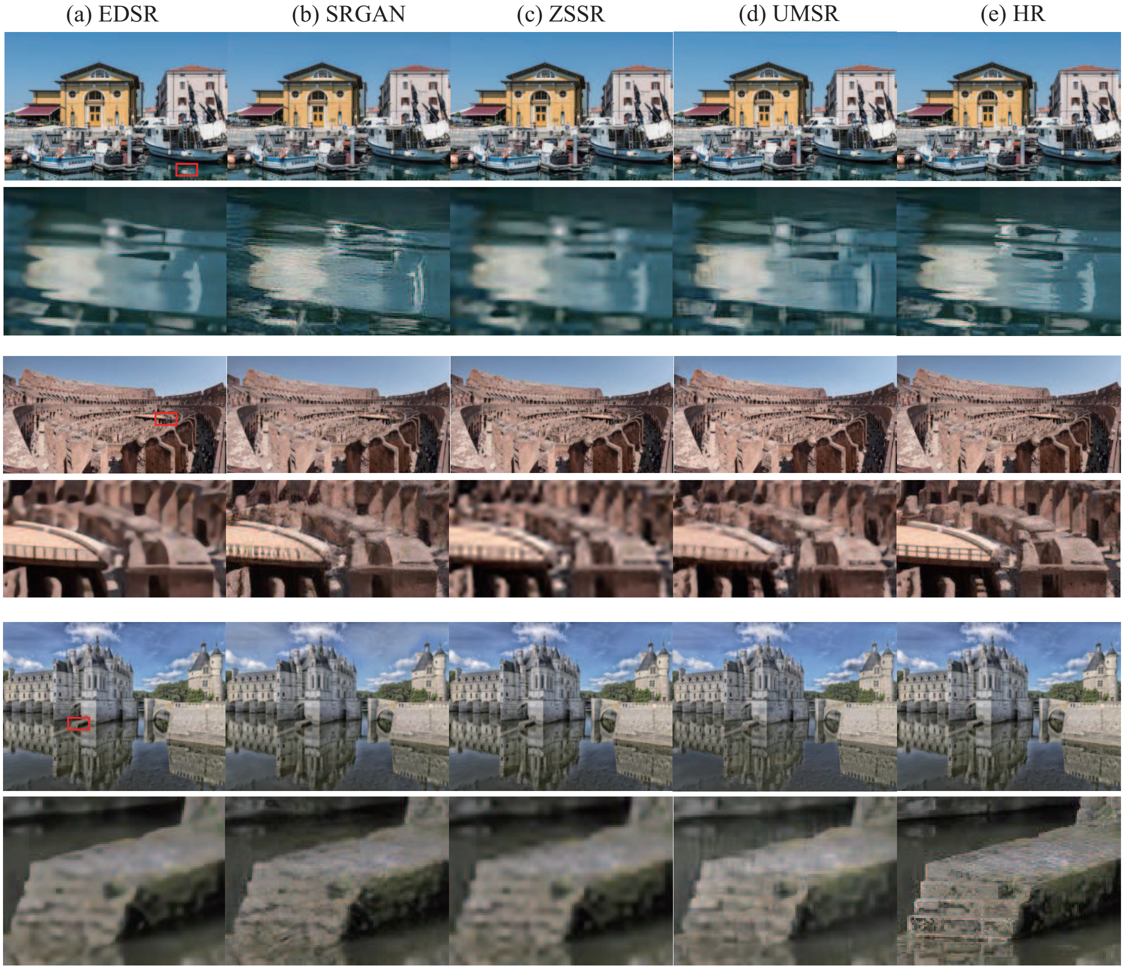


Figure 7. Comparison on supervised and unsupervised methods. (**a**)EDSR; (**b**)SRGAN; (**c**)ZSSR; (**d**)UMSR; (**e**)HR.

Table 1. Comparison on different methods with DIV2K dataset. The image is 4× scaled and no cropping is used during the test period. We report the PSNR scores.

PSNR	EDSR	ZSSR	SRGAN	UMGSR (MSE)	UMGSR (MSE + Percp)	UMGSR (Total Loss)
Image1	27.74	24.72	24.05	25.05	25.02	24.89
Image2	25.03	23.81	22.83	23.96	24.03	23.87
Image3	27.45	26.74	24.46	24.78	24.93	24.87

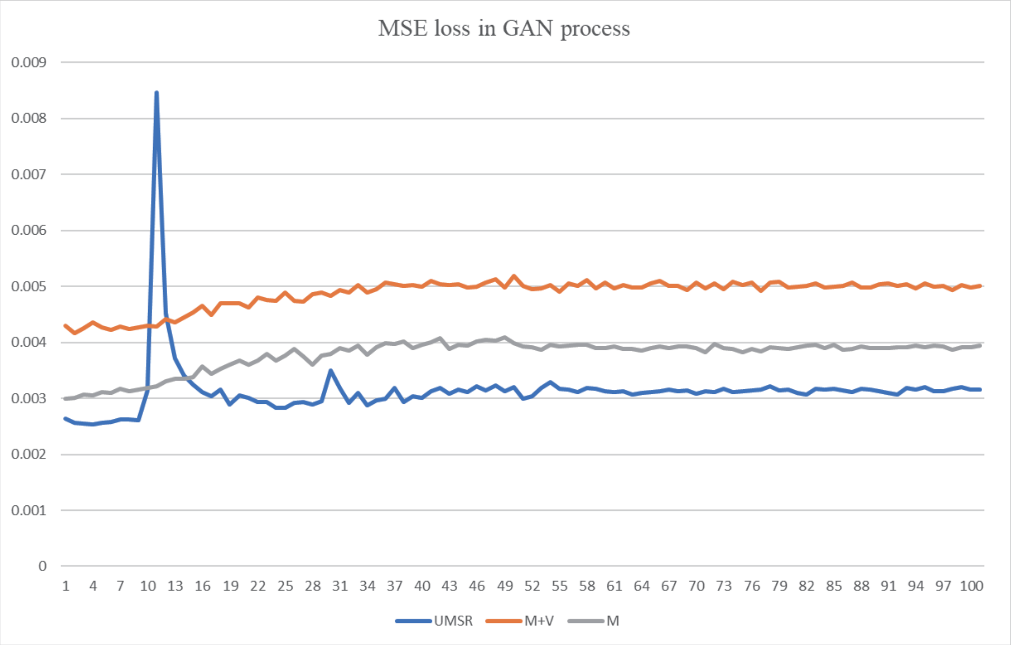


Figure 8. Comparison the MSE error of results with MSE loss (M), MSE and perceptual loss (M+V), and MSE, Perceptual loss and Multi-gram loss (UMGSR).

4.3. Discussion

In this paper, we compare the proposed UMGSR with other state-of-the-art supervised and unsupervised methods with both traditional PSNR value and the power-spectrum image contrast. Referring to the unsupervised setting of UMGSR, more analysis needs to be involved, to better evaluate its performance. On the other hand, the latest research in [50] suggests that there is a trade-off between distortion and perception. Our research pays much attention to the visual satisfactory generation, which hurts the PSNR to some extent. Hence, traditional accuracy measurement, such as MSE, PSNR, and SSIM [51] cannot justify the advantage of our method properly.

We exhibit the SR results of five different methods, EDSR, ZSSR, SRGAN, UMGSR (MSE), and UMGSR (total loss), with HR images in Figure 9. The PSNR scores are shown in Table 1. In detail, image 1 is from DIV2K [49]. It acts as the training image of EDSR. According to the PSNR values, EDSR achieves the best result.

On the other hand, from Figure 7, we can infer that UMGSR produces SR image with more carving details, leading to better visual feeling than EDSR. The conclusion is in keeping with the viewpoint of SRGAN: higher PSNR does not guarantee a better perceptual result. This phenomenon is fairly obvious in the comparison between UMGSR with MSE loss and with total loss. In unsupervised SR learning, PSNR of ZSSR is much higher than ours while their SR images are in worse visual details. To highlight the difference among these methods, we compare the SR images by their 3D power-spectrum [52] in Figure 9. From the spectrum distribution, we can clearly see the distribution of the whole image. It distinctly shows that our method is much better than ZSSR and EDSR, which generate obvious faults. We assume that it is due to the mixture loss leading to better texture generation ability in our model.

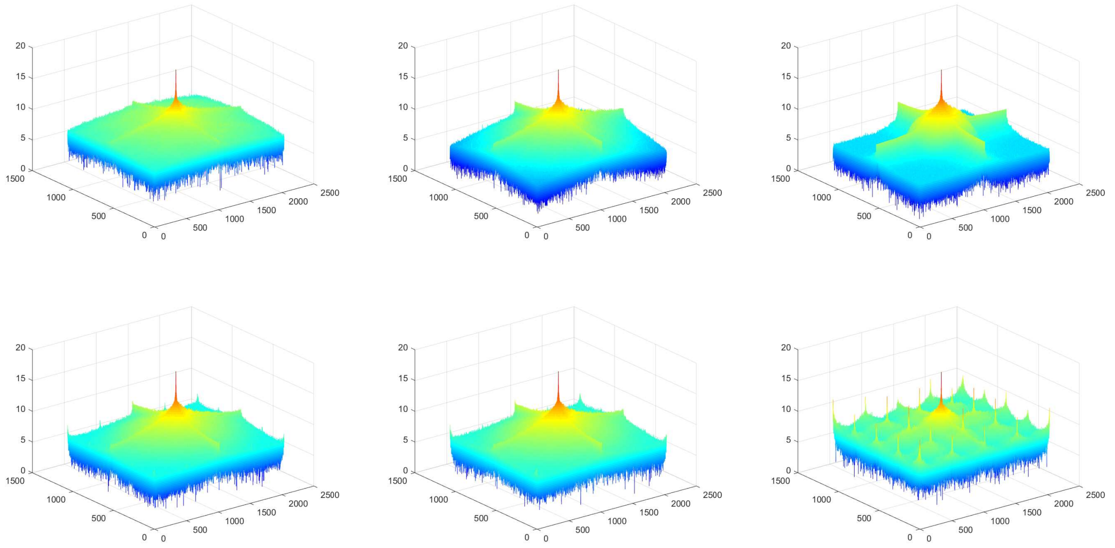


Figure 9. The power spectra of the second image in Figure 7: HR, EDSR, ZSSR, UMGSR with MSE loss, and UMGSR with total loss. Smooth edge of spectra reflects more colorful details and sharp fault means the lack of some color range. Even though abundant power spectra does not mean accurate, it indeed prove more vivid details in the image. As a result, our model can generate dramatic features than accurate pursuing models(EDSR, ZSSR).

To better evaluate these models, we show generations in the same chosen patch in Figure 7. These results show that traditional accuracy-pursuing SR methods generate rough details and better shape lines, while UMGSR (total loss) results in satisfactory performance in image details, which are even better than the supervised SRGAN. This is also verified in 3D power-spectrum image, where our result is quite similar to the HR.

In general, high-frequency information (like shape lines) is more sensitive to accuracy driven methods, such as EDSR. Meanwhile, SR images generated by these methods hardly provide pleasant visual feeling. Their ensembles are like drawn or cartoon images. For example, Roma Desert place (The second test image -3*rd*and 4*th* rows in Figure 7) generated by EDSR shows sharper edges but untrue effect. Visual feeling pursuing models (like SRGAN and UMGSR) generate more photo-realistic features accompanied by inaccurate information in pixel-level. For example, SRGAN introduces rough details in the local parts far away from the ground truth, especially for the large flat space. In our opinion, this is the common weakness of GAN related SR methods. In particular, our two-step learning partly overcomes it. Accordingly, the SR images of UMGSR show better shapes than SRGAN along with better visual feeling than EDSR.

5. Conclusions and Future Work

In this paper, we propose a new unsupervised SR method: UMGSR, for the scenario of no supervised HR image involved. Compared with former supervised and unsupervised SR methods, UMGSR mainly introduces both a novel architecture and a new multi-gram loss. With these modifications, our UMGSR can address SR issue with single input in any condition. Experimental results show that UMGSR can generate better texture details than other unsupervised methods. In the future work, we will pay more attention to combining our model with GANs on supervised SR problems.

Author Contributions: Project administration, funding acquisition, guidance for the research, and revised the paper, Y.S.;writing–original draft preparation, data curation, software, methodology, writing–review and editing and supervision, B.L.; supervision, writing–review and editing, and funding acquisition, B.W.; guidance for the research, conceptualization, software, validation, and supervision, Z.Q.; visualization, supervision, J.L..

Funding: This research was funded by the National Natural Science Foundation of China No. 91546201, No. 71331005, No. 71501175, and No. 61702099, and the Open Project of Key Laboratory of Big Data Mining and Knowledge Management, Chinese Academy of Sciences, and the Foundation for Disciplinary Development of SITM in UIBE.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Lim, B.; Son, S.; Kim, H.; Nah, S.; Mu, Lee, K. Enhanced deep residual networks for single image super-resolution. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, Honolulu, HI, USA, 21–26 July 2017; pp. 136–144.
2. Haris, M.; Shakhnarovich, G.; Ukita, N. Deep Back-Projection Networks For Super-Resolution. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 1664–1673.
3. Ledig, C.; Theis, L.; Huszár, F.; Caballero, J.; Cunningham, A.; Acosta, A.; Aitken, A.; Tejani, A.; Totz, J.; Wang, Z.; et al. Photo-Realistic Single Image Super-Resolution Using a Generative Adversarial Network. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 4681–4690.
4. Dong, C.; Loy, C.C.; He, K.; Tang, X. Learning a Deep Convolutional Network for Image Super-Resolution. In *European Conference on Computer Vision*; Springer: Cham, Switzerland, 2014; pp. 184–199.
5. Kim, J.; Kwon, Lee, J.; Mu, Lee, K. Accurate Image Super-Resolution Using Very Deep Convolutional Networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 26 June–1 July 2016; pp. 1646–1654.
6. Tai, Y.; Yang, J.; Liu, X. Image Super-Resolution via Deep Recursive Residual Network. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 3147–3155.
7. Caballero, J.; Ledig, C.; Aitken, A.; Acosta, A.; Totz, J.; Wang, Z.; Shi, W. Real-Time Video Super-Resolution with Spatio-Temporal Networks and Motion Compensation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 4778–4787.
8. Sajjadi, M.S.M.; Scholkopf, B.; Hirsch, M. EnhanceNet: Single Image Super-Resolution Through Automated Texture Synthesis. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 4501–4510.
9. Shi, W.; Caballero, J.; Ledig, C.; Zhuang, X.; Bai, W.; Bhatia, K.; de Marvao, A.M.S.M.; Dawes, T.; O'Regan, D.; Rueckert, D. Cardiac image super-resolution with global correspondence using multi-atlas patchmatch. In Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention, Nagoya, Japan, 22–26 September 2013; pp. 9–16.
10. Zou, W.W.W.; Yuen, P.C. Very Low Resolution Face Recognition Problem. *IEEE Trans. Image Process.* **2011**, *21*, 327–340. [CrossRef] [PubMed]
11. Shocher, A.; Cohen, N.; Irani, M. "Zero-Shot" Super-Resolution using Deep Internal Learning. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 3118–3126.
12. Mechrez, R.; Talmi, I.; Zelnik-Manor, L. Image super-resolution based on local self-similarity. In Proceedings of the European Conference on Computer Vision, Munich, Germany, 8–14 September 2018; pp. 768–783.
13. Gatys, L.A.; Ecker, A.S.; Bethge, M. A Neural Algorithm of Artistic Style. *arXiv* **2015**, arXiv:1508.06576.
14. Gatys, L.; Ecker, A.S.; Bethge, M. Texture Synthesis Using Convolutional Neural Networks. In *Advances in Neural Information Processing Systems*; NIPS: Grenada, Spain, 2015; pp. 262–270.
15. Blau, Y.; Mechrez, R.; Timofte, R.; Michaeli, T.; Zelnik-Manor, L. 2018 PIRM Challenge on Perceptual Image Super-resolution. In Proceedings of the European Conference on Computer Vision, Munich, Germany, 8–14 September 2018.
16. Zhang, Y.; Li, K.; Li, K.; Wang, L.; Zhong, B.; Fu, Y. Image Super-Resolution Using Very Deep Residual Channel Attention Networks. In Proceedings of the European Conference on Computer Vision, Munich, Germany, 8–14 September 2018; pp. 286–301.

17. Zhang, Y.; Tian, Y.; Kong, Y.; Zhong, B.; Fu, Y. Residual Dense Network for Image Super-Resolution. In Proceedings of the IEEE International Conference on Computer Vision, istanbul, Turkey, 30–31 January 2018; pp. 2472–2481.
18. Zhao, H.; Gallo, O.; Frosio, I.; Kautz, J. Loss Functions for Image Restoration With Neural Networks. *IEEE Trans. Comput. Imaging* **2016**, *3*, 47–57. [CrossRef]
19. Bulat, A.; Tzimiropoulos, G. Super-FAN: Integrated facial landmark localization and super-resolution of real-world low resolution faces in arbitrary poses with GANs. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 109–117.
20. Irani, D.G.S.B.M. Super-resolution from a single image. In Proceedings of the IEEE International Conference on Computer Vision, Kyoto, Japan, 29 September–2 October 2009; pp. 349–356.
21. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 26 June–1 July 2016; pp. 770–778.
22. Snelgrove, X. High-resolution multi-scale neural texture synthesis. In *SIGGRAPH Asia Technical Briefs*; ACM SIGGRAPH: Los Angeles, CA, USA, 2017; pp. 1–13.
23. Chang, H.; Yeung, D.-Y.; Xiong, Y. Very Low Resolution Face Recognition Problem. In Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Washington, DC, USA, 27 June–2 July 2004.
24. Duchon, C.E. Lanczos Filtering in One and Two Dimensions. *J. Appl. Meteorol.* **1979**, *18*, 1016–1022. [CrossRef]
25. Thornton, M.W.; Atkinson, P.M.; Holland, D.A. Sub-pixel mapping of rural land cover objects from fine spatial resolution satellite sensor imagery using super-resolution pixel-swapping. *Int. J. Remote. Sens.* **2006**, *27*, 473–491. [CrossRef]
26. Ulyanov, D.; Vedaldi, A.; Lempitsky, V. Deep Image Prior. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 9446–9454.
27. Protter, M.; Elad, M. Super Resolution With Probabilistic Motion Estimation. *IEEE Trans. Image Process.* **2009**, *18*, 1899–1904. [CrossRef] [PubMed]
28. Suetake, N.; Sakano, M.; Uchino, E. Image super-resolution based on local self-similarity. *Opt. Rev.* **2008**, *15*, 26–30. [CrossRef]
29. Lin, Z.; Shum, H.-Y. Fundamental Limits of Reconstruction-Based Superresolution Algorithms under Local Translation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2004**, *26*, 83–97. [PubMed]
30. Redmon, J.; Farhadi, A. YOLO9000: Better, faster, stronger. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 7263–7271.
31. Chen, L.C.; Papandreou, G.; Kokkinos, I.; Murphy, K.; Yuille, A.L. DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs. *IEEE Trans. Pattern Anal. Mach. Intell.* **2018**, *40*, 834–848. [CrossRef] [PubMed]
32. Sandler, M.; Howard, A.; Zhu, M.; Zhmoginov, A.; Chen, L.-C. Mobilenetv2: Inverted residuals and linear bottlenecks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 4510–4520.
33. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. ImageNet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*; NIPS: Grenada, Spain, 2012; pp. 1097–1105.
34. Dong, C.; Loy, C.C.; He, K.; Tang, X. Image Super-Resolution Using Deep Convolutional Networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2015**, *38*, 295–307. [CrossRef] [PubMed]
35. Kim, J.; Kwon Lee, J.; Mu Lee, K. Deeply-Recursive Convolutional Network for Image Super-Resolution. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 26 June–1 July 2016; pp. 1637–1645.
36. Tong, T.; Li, G.; Liu, X.; Gao, Q. Image Super-Resolution Using Dense Skip Connections. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 4799–4807.
37. Huang, G.; Liu, Z.; Van Der Maaten, L.; Weinberger, K.Q. Densely Connected Convolutional Networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Minneapolis, MN, USA, 17–22 June 2007; pp. 4700–4708.
38. Capel, D.L. Image mosaicing. In *Image Mosaicing and Super-Resolution*; University of Oxford: Oxford, UK, 2004; pp. 47–79.

39. Farsiu, S.; Robinson, M.D.; Elad, M.; Milanfar, P. Fast and Robust Multiframe Super Resolution. *IEEE Trans. Image Process.* **2004**, *13*, 1327–1344. [CrossRef] [PubMed]
40. Baker, S.; Kanade, T. Limits on super-resolution and how to break them. *IEEE Trans. Pattern Anal. Mach. Intell.* **2002**, *9*, 1167–1183. [CrossRef]
41. Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; Bengio, Y. Generative Adversarial Nets. In *Advances in Neural Information Processing Systems*; NIPS: Grenada, Spain, 2014; pp. 2672–2680.
42. Karras, T.; Aila, T.; Laine, S.; Lehtinen, J. Progressive Growing of GANs for Improved Quality, Stability, and Variation. *arXiv* **2017**, arXiv:1710.10196.
43. Johnson, J.; Alahi, A.; Fei-Fei, L. Perceptual losses for real-time style transfer and super-resolution. In Proceedings of the European Conference on Computer Vision, Amsterdam, The Netherlands, 11–14 October 2016; pp. 694–711.
44. Freedman, G.; Fattal, R. Image and video upscaling from local self-examples. *ACM Trans. Graph. (TOG)* **2011**, *30*, 12. [CrossRef]
45. Huang, J.-B.; Singh, A.; Ahuja, N. Single image super-resolution from transformed self-exemplars. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 5197–5206.
46. Zontak, M.; Irani, M. Internal statistics of a single natural image. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Colorado Springs, CO, USA, 20–25 June 2011; pp. 977–984.
47. Simonyan, K.; Zisserman, A. Very Deep Convolutional Networks for Large-Scale Image Recognition. *arXiv* 2014, arXiv:1409.1556.
48. Zeyde, R.; Elad, M.; Protter, M. Texture Synthesis Using Convolutional Neural Networks. In Proceedings of the International Conference on Curves and Surfaces, Avignon, France, 24–30 June 2010; pp. 711–730.
49. Timofte, R.; Agustsson, E.; Van Gool, L.; Yang, M.-H.; Zhang, L. NTIRE 2017 Challenge on Single Image Super-Resolution: Methods and Results. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshopsh, Honolulu, HI, USA, 21–26 July 2017; pp. 114–125.
50. Blau, Y.; Michaeli, T. The Perception-Distortion Tradeoff. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 6228–6237.
51. Wang, Z.; Bovik, A.C.; Sheikh, H.R.; Simoncelli, E.P. Image quality assessment: From error visibility to structural similarity. *IEEE Trans. Image Process.* **2004**, *13*, 600–612. [CrossRef] [PubMed]
52. Van der Schaaf, V.A.; van Hateren, J.H.V. Modelling the power spectra of natural images: Statistics and information. *Vis. Res.* **1996**, *36*, 2759–2770. [CrossRef]