remote sensing

MDPI

Article

Unpaired Remote Sensing Image Super-Resolution with Multi-Stage Aggregation Networks

Lize Zhang [1], Wen Lu [1,*], Yuanfei Huang [2], Xiaopeng Sun [1] and Hongyi Zhang [1]

[1] School of Electronic Engineering, Xidian Universerty, Xi'an 710071, China; lzzhang_98@stu.xidian.edu.cn (L.Z.); xpsun@stu.xidian.edu.cn (X.S.); hyzhang_12@stu.xidian.edu.cn (H.Z.)
[2] School of Artificial Intelligence, Beijing Normal University, Beijing 100875, China; yfhuang@bnu.edu.cn
* Correspondence: luwen@mail.xidian.edu.cn

Abstract: Mainstream image super-resolution (SR) methods are generally based on paired training samples. As the high-resolution (HR) remote sensing images are difficult to collect with a limited imaging device, most of the existing remote sensing super-resolution methods try to down-sample the collected original images to generate an auxiliary low-resolution (LR) image and form a paired pseudo HR-LR dataset for training. However, the distribution of the generated LR images is generally inconsistent with the real images due to the limitation of remote sensing imaging devices. In this paper, we propose a perceptually unpaired super-resolution method by constructing a multi-stage aggregation network (MSAN). The optimization of the network depends on consistency losses. In particular, the first phase is to preserve the contents of the super-resolved results, by constraining the content consistency between the down-scaled SR results and the low-quality low-resolution inputs. The second stage minimizes perceptual feature loss between the current result and LR input to constrain perceptual-content consistency. The final phase employs the generative adversarial network (GAN) to adding photo-realistic textures by constraining perceptual-distribution consistency. Numerous experiments on synthetic remote sensing datasets and real remote sensing images show that our method obtains more plausible results than other SR methods quantitatively and qualitatively. The PSNR of our network is 0.06dB higher than the SOTA method—HAN on the UC Merced test set with complex degradation.

Keywords: remote sensing; unpaired super-resolution; multi-stage aggregation network; consistency losses

Citation: Zhang, L.; Lu, W.; Huang, Y.; Sun, X.; Zhang, H. Unpaired Remote Sensing Image Super-Resolution with Multi-Stage Aggregation Networks. *Remote Sens.* **2021**, *13*, 3167. https://doi.org/10.3390/rs13163167

Academic Editors: Karen Egiazarian, Vladimir Lukin and Aleksandra Pizurica

Received: 15 June 2021
Accepted: 6 August 2021
Published: 10 August 2021

1. Introduction

Image super-resolution, which aims to reconstruct the high-resolution (HR) image from its low-resolution (LR) observation, is an active research topic and has been demonstrated to be an effective method to increase the spatial resolution. In the field of remote sensing, the development of antenna arrays [1,2] plays an essential role in image super-resolution. SR methods compensate for the information lost in the process of image transmission and compression, and improve the spatial resolution of remote sensing data for environmental monitoring [3] and object detection. The remote sensing image degradation process is usually defined as

$$I_{LR} = (I_{HR} \bigotimes k) \downarrow_s + n, \tag{1}$$

where I_{HR} means high-resolution images, k denotes blur kernels, $\downarrow_s$ means degradation model with scale factor s, n is additive white Gaussian noise(AWGN) in general. Comparing with updating the hardware devices, using the SR technique has advantages of low cost, easy implementation, and high efficiency. Remote sensing image super-resolution has become one of the most important applications of SR technology.

Previous super-resolution methods are mainly based on interpolation [4] and reconstruction [5]. Images with different spatial resolutions have different carrying capacities for information. The problem of image super-resolution is from reconstruction of LR images with severe lack of information to HR images with rich information is an ill-posed problem that a single LR image can reconstruct multiple super-resolved results. To tackle this problem, the example-learning based SR methods have been proposed to learn the mapping function between LR and HR image pair, such as sparse coding [6,7] and probabilistic graphical model [8]. Furthermore, with the development of deep convolutional neural network(CNN) in the field of computer vision, the SR methods based on deep learning came into being because they have a higher capability of representation for LR-to-HR nonlinear mapping modeling than the previous example-learning based SR methods. As an initial attempt, Dong et al. [9] introduced CNN into SR that employed three convolutional layers to model the mapping function. In order to increase the depth of the network for improving the capability of the model, Kim et al. [10] firstly introduced the global residual learning and gradient clipping strategy into the SR task and succeeds in training a deep SR network by using global residual learning and gradient clipping strategy. Inspired by VDSR, Tai et al. [11] used local residual learning and global residual learning to build a deep recursive residual network to improve the effect of super-resolved results. Skip connection based networks [12,13] can fuse information in different network depths to increase the representation ability of the model so that it has a stronger information-carrying capacity. Tong et al. [14] initially employed densely connected block and proposed SR method based on dense skip connections. EDSR [15] which introduced local residual learning and removed unnecessary batch normalization layers made a significant breakthrough in SR performance. Ledig et al. [16] applied Generative Adversarial Networks [17] into the model and proposed the perceptual loss function. It is crucial to generate more realistic textures and images. ESRGAN [18] employed residual in residual dense blocks and optimized loss function on SRGAN. Zhang et al. [19] integrated the channel-wise attention into the residual blocks by recalibrating the feature channel to push the state-of-the-art(SOTA) performance of SR forward. Huang et al. [20] initially proposed an interpretable attention mechanism into SR task and achieved SOTA performance. Supervised methods usually use a particular degradation model from HR as LR to train the network. Some researches [21,22] have shown that critical to the success of an accurate model of the degradation of SISR. However, in contrast to the natural images dataset, the original image in remote sensing dataset [23] is not clear enough to meet the need as a ground-truth for training. In addition, paired HR remote sensing images are hard to collect, especially since the cost of updating hardware devices is extremely high.

However, the current mainstream SR methods focus on model innovation, on the one hand, assuming that the image degradation process is bicubic interpolation. On the other hand, focus on the degradation process and build models to fit the data under various blur kernels. When LR images with unknown degradation that real remote sensing images contain the sensor noise and aliasing effects, the existing SR methods struggle to achieve satisfactory performance since we don't know the downsampling operation of remote sensing images. Nevertheless, unpaired super-resolution method has a good performance in constructing real natural images. Freedman et al. [24] proposed self-similarity based image SR, which reconstructed the HR images by using patches in previous natural images. Yuan et al. [25] proposed cycle consistency and designed Cycle-in-Cycle GAN, which did not require the paired LR-HR images and solved the unsupervised image translation problems. First, a GAN with two discriminators was employed for learning degradation process; then another GAN was used for image SR. Shocher et al. [26] exploited the internal self-similarity of a single image internal information into self-supervised iterative optimization of the model without any reference images or pretrained. Ulyanov et al. [27] proposed a new strategy for dealing with the regularization tasks in inverse problems and has been proved to be very effective and has been successful in many imaging inverse problems such as denoising, super-resolution and so on. Since the imaging environment

of remote sensing images is generally unknown and complex, unsupervised learning has been developed in remote sensing image SR. Zhang et al. [28] proposed an unsupervised method that employs GAN to obtain super-resolved and achieved satisfactory performance. Inspired by CycleGAN, Wang et al. [29] proposed two generative CNNs for down-sampling named Cycle-CNN modeling the degradation process and the corresponding reconstruction, named CycleCNN. UDGN [3] models the internal recurrence of information inside the single image to generate Lake Area images with higher spatial resolution without requiring pretraining. Zhang et al. [30] proposed a multi-degradation aided method by adopting the multiple Gaussian blur kernels and AWGN to get LR images for adapting to the mixed degraded model.

Since there is not any prior degradation, they generate bad content and details that do not match the corresponding LR image, many unpaired super-resolution methods can only use image features that rely on LR patches of a small fixed size. The spatial distribution of remote sensing images is highly complex, so reconstructing a high-quality HR image with accurate content and refining photo-realistic details are necessary. When we get wrong content from low-quality remote sensing images, we observe that super-resolved results produced by GAN-based methods are often affected by structural distortions. To tackle this problem and inspired by DCSR [31], we explicitly add a multi-stage aggregation architecture which is a step-by-step training mode to realize content consistency(CC), perceptual-content consistency(PCC) and perceptual-distribution consistency(PDC). We also propose consistency losses for our network. The contributions of our proposed unpaired SR method can be summarized in three points:

1. We introduce a multi-stage aggregation network for gradually optimizing the model with the degraded self-exemplars and unpaired references, which allows it to achieve effective optimization from content to perception. Specifically, the first stage can be adapted to better pixel-wise PSNR, and the subsequent stages can be adapted to more realistic texture and details reconstruction.
2. Aiming at retaining the content on remote sensing images and excavating its underlying perceptual similarities from the low-quality images, we propose consistency loss functions for contents retainment and details reconstruction in different phases.
3. We conducted experimental validation of multiple datasets on our method, and the results indicate the superiority of our method in remote sensing SR and have more intuitive visual effects.

2. Materials and Methods

2.1. Methods

The SR task aims to estimate the SR image I_{SR} from its LR counterpart I_{LR}. Generally, the dataset of paired super-resolution tasks has LR-HR paired images and LR image input model to receive SR image by non-linear mapping. While there are typically no adequate paired high-quality remote sensing images as training reference so that unpaired SR method has received more and more attention. In most cases, we get LR images from original images assuming that the image degradation process is the bicubic interpolation. But in fact, we only have low-quality low-resolution images as input so that reconstructing photo-realistic details is challenging. Therefore, our proposed method aims at recovering more desired details of LR images with unpaired training datasets.

The framework of MSAN which is shown in Figure 1 consists of a generator and a discriminator and shows the process of calculating consistency loss functions. In the different stages, the generator can be updated by calculating corresponding consistency losses. Given an input LR image I_L without paired HR image. The super-resolved image I_{SR} can be generated by

$$\begin{aligned} I_{SR_C} &= H_{Generator-C}(I_L)_r \\ I_{SR_{PC}} &= H_{Generator-PC}(I_L)_r \\ I_{SR} &= H_{Generator-PD}(I_L)_r \end{aligned} \tag{2}$$

where, $H_{Generator-C}(\cdot)$, $H_{Generator-PC}(\cdot)$ and $H_{Generator-PC}(\cdot)$ denote the generator in different stages respectively, and r is the scale factor that we want to LR enlarge. And the structure of Generator-C, Generator-PC and Generator-PD are introduced in detail in Figure 2. For extracting and aggregating the multi-level features, we design an efficient framework which contains dense connections and skip connections on multi-level fusion module. Furthermore, in order to achieve the photo-realistic performance, we utilize multi-stage consistency for retaining content and structure from low-resolution inputs and recovering more photo-realistic details. In this section, we detail our proposed method MSAN.

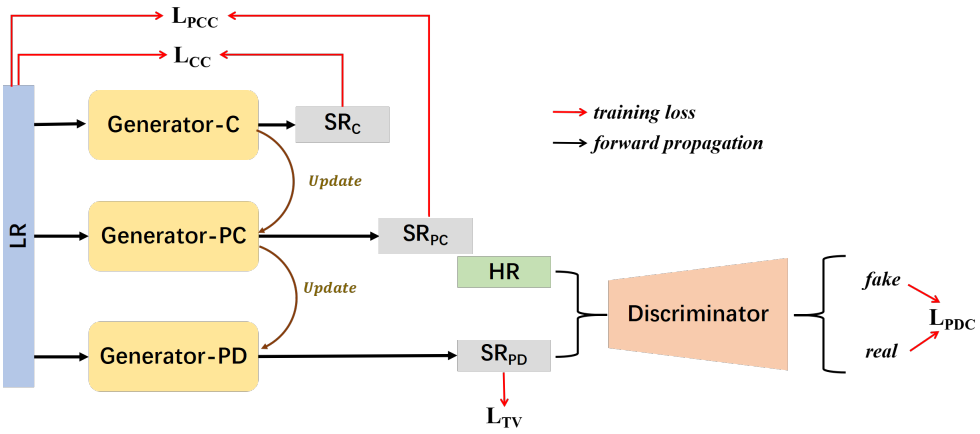

Figure 1. Overview of the proposed unpaired SR method.

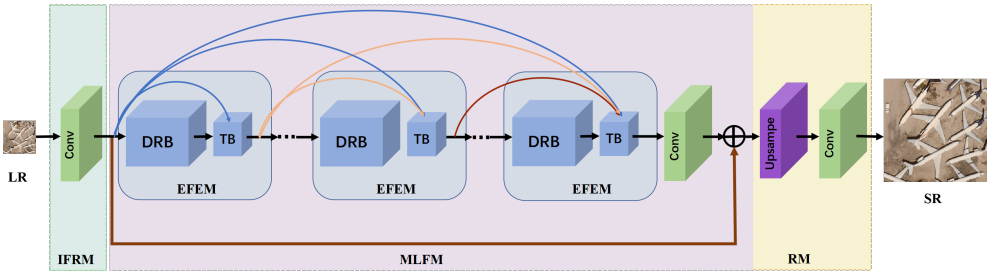

Figure 2. The framework of generator.

2.1.1. Network Architecture

For training unpaired super-resolution network, we don't have a training set $\{I_{LR_i}, I_{HR_i}\}^{N_{i=1}}$, that has N LR-HR pairs. Moreover, remote sensing images generally incorporate various degradation models, including poor details and noises that are undesired for the image SR task. However, noise and structure degradation models are important for recovering principle contents and un-occurred details; therefore, there is an urgent need to design a module that effectively utilizes low-quality input to extract low-frequency image feature information.

As shown in Figure 2, following our previous work DCSR [31], the generator mainly consists of three parts: initial feature representation module (IFRM), multi-level fusion module (MLFM) and reconstruction module (RM). The IFRM is applied to represent a three-channel LR input as a set of feature-maps, similar to the most existing super-resolution methods. The MLFM is designed to capture more informative features for SR by a sequence of stacked enhanced feature extraction modules (EFEM). The extracted features are fed into the RM to reconstruct the SR images. We employ a convolutional layer as the IFRM to extract the initial features from LR input, which can be empressed by

$$f^0 = H_{IFRM}(I_L) = Conv_3(I_L) \tag{3}$$

$$H_{generator}(\cdot) = H_{IFRM}(\cdot) + H_{MLFM}(H_{IFRM}(\cdot)) \tag{4}$$

where, f_0 is the initial extracted feature which is then fed into the MLFM and also used for global residual-feature learning, $H_{IFRM}(\cdot)$ and $H_{DFE}(\cdot)$ respectively denote the function of initial feature representation and multi-level fusion module.

Inspired by the dense connections [32] and residual blocks [12,13], they have been expounded to be an effective strategy for benefiting the information flow of transferred features at different layers. Different from RDN [13], we add an adaptive learning filter after concatenation. It is worth mentioning that the main contribution of our proposed work is a multi-stage aggregation strategy so that the generator of our network follows the existing networks and can be replaced with any better backbone. The basic dense network is illustrated in Figure 2. We introduce the strategies in our multi-level fusion module. For super-resolution task, shallow-layer features generally represent the coarse image information, however, deep-layer feature maps represent the fine details and textures. Therefore, in order to improve the reuse of features in the middle layer, we employ the local skip-connections to ensure that the shallow-layer information still participates in the image reconstruction. The MLFM is constructed by stacking n enhanced feature extraction modules (EFEM) for inferring the informative features for detail recovering as shown in Figure 2. Let f_{dense} represent the feature generated by the dense-sampling structure. Thus, we introduce the MLFM as

$$f_{dense} = Conv_3(H^n_{EFEM}(H^{n-1}_{EFEM}(H^{n-2}_{EFEM}(\ldots(H^1_{EFEM}(f^0))\ldots)))) + f^0 \tag{5}$$

where $H^n_{EFEM}(\ldots)$ indicates the composite function of n-th EFEM, $Conv_3(\cdot)$ indicates a 3×3 convolution operation.

As illustrated in Figure 3, different from Feature Extraction Module (FEM) in DCSR, we add a transition block (TB) which can combine information coming from different blocks through an adaptive learning process. We also improve the previous FEM that employs dense connections between m residual blocks (DRB) to form a residual blockchain. To improve the continuous memory of information flow, TB is attached to the fusing of feature information of each layer better. In the transition block, The features extracted by the previous EFEMs and the current residual blockchain are concatenated and then input a convolutional layer to process features more efficiently. Specifically, the inputs $f_0, f_1, \ldots,$ and f_i are subjected to the concatenation operation. Let f_n represent the feature generated by the middle layer. The process of DRB and TB are formulated as

$$x^m = r_m([x^0, x^1, \ldots, x^{m-1}]) \tag{6}$$

where x^m denotes the feature extracted by the m-th residual block, $[x^0, x^1, x^2, \ldots, x^{m-1}]$ indicate the features are concatenated together and r_0 denotes

$$r_0(f^{n-1}) = Conv_3(ReLU(Conv_3(f^{n-1}))) + f^{n-1} \tag{7}$$

$$H^n_{DRB} = r_m(r_{m-1}(\ldots r_0(f_{n-1}))) \tag{8}$$

$$f_n = H^n_{EFEM}(f_{n-1}) = \varphi_n([H^n_{DRB}(f_{n-1}), f_{n-1}, f_{n-2}, \ldots, f_0]) \tag{9}$$

where $\varphi_n(\cdot)$ denotes the function of a convolutional layer that kernel size is one which is applied for reducing the channel dimension into a fixed size, $H^n_{DRB}(\cdot)$ denotes the dense residual blockchain contained in the n-th EFEM.

After acquiring the high informative features, we utilize an representation module (RM) for feature manipulation and outputting a high-resolution image. As shown in Figure 2 the final representation module only consists of one learnable layer which is a convolution operation and a non-parametric operation (sub-pixel convolution [33]). When we need to enlarge the image on scale factor s, RM rearrange the tensor with dimensions $H \times W \times C \cdot r^2$ as $rH \times rW \times C$ for converting groups of high-resolution features into RGB images.

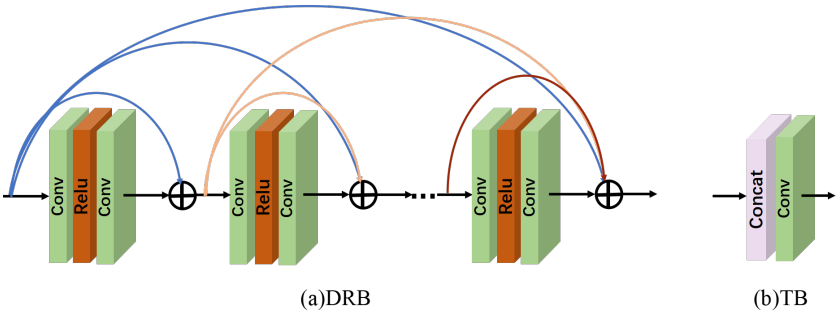

Figure 3. The architecture of DRB (**a**) and TB (**b**).

2.1.2. Multi-Stage Architecture

Traditionally, we can optimize the model parameters to achieve good performance by minimizing the loss function between the super-resolved results and ground-truth, as shown in Equation (10):

$$min \sum_{n=1}^{N} Loss_n(I_{SR}, I_{HR}) \tag{10}$$

For the unpaired remote sensing super-resolution task, the low-resolution remote sensing images still contain enough similar principal information by transmitting hardware devices. Therefore, multi-stage aggregation can solve the problem that unpaired methods generate undesirable content and details, and consistency losses are introduced to restrict model convergence, which is shown in Figure 1.

(1) First stage: Content Consistency

Aiming at recovering the principal information of the super-resolved results from the LR inputs, the content-consistency stage is optimized via minimizing the difference between the super-resolved image I_{SR_C} and the corresponding low-quality image I_L. We proposed a content-consistency loss function (L_{CC}) in the training process of the first stage. Only using L_{CC} to optimize the model during training can reconstruct the correct content and avoid recovering the wrong content due to the high spatial resolution of the remote sensing images. In this stage, the content-consistency loss L_{CC} is formulated as

$$L_{CC} = \left\| \left(D(I_{SR_C}) - I_L \right) \right\|_1 \tag{11}$$

where $D(\cdot)$ indicates the down-scaling operators using the bicubic algorithm. We down-sample super-resolved images I_{SR_C} by the same scale factor and the processed image as input performs enhanced feature extraction and reconstruction. This stage mainly pursues a high PSNR value. However, due to the influence of uncontrollable artifacts under the unknown degradation model, noise and distortion will still affect the reconstruction of the principal information at this stage. The super-resolved of the current stage is slightly smoother and cannot produce rich details.

(2) Second stage: Perceptual-Content Consistency

In the absence of paired groud-truth images, while ensuring that the content remains consistent, excavating deep semantic features is urgently necessary. In order to enhance the perceptual content of super-resolved images and explore the implicit perceptual similarity, we perform the second stage of training on the results of the previous stage. Particularly, we design two consistency loss functions (L_{PCC_1} and L_{PCC_2}) as illustrated in Figure 4. The generator used at this stage is the same as in the previous stage. The pre-trained VGG-19 [34] is applied here for inferring the perceptual similarity of I_L and $I_{SR_{PC}}$. As shown in Figure 4, in calculating L_{PCC_1}, we perform a blurring operation on $I_{SR_{PC}}$ and upscaled I_L. The objective functions for training the second phase could be formulated as

$$L_{PCC_1} = \left\| \left(\varphi_{deep}(G(I_{SR_{PC}})) - \varphi_{deep}(G(Up(I_L))) \right) \right\|_1 \tag{12}$$

where $Up(\cdot)$ indicates the up-scaling operators via bicubic resampling meanwhile the size of $Up(I_L)$ is the same as $I_{SR_{PC}}$. Aiming at removing the distortion effects, we apply the blurring operators $G(\cdot)$ on images using 5×5 Gaussian blur kernels with $\sigma = 1.2$. Furthermore, we use the deep layer features φ_{deep} of VGG-19 for representing the perceptual-content (PC) representations. We extract the final perceptual-content features on the "$Conv_{5,4}$" layer of VGG19. Specifically, "$Conv_{m,n}$" indicates the feature maps obtained by the n-th convolution (before activation) before the m-th max-pooling layer within the VGG-19 network.

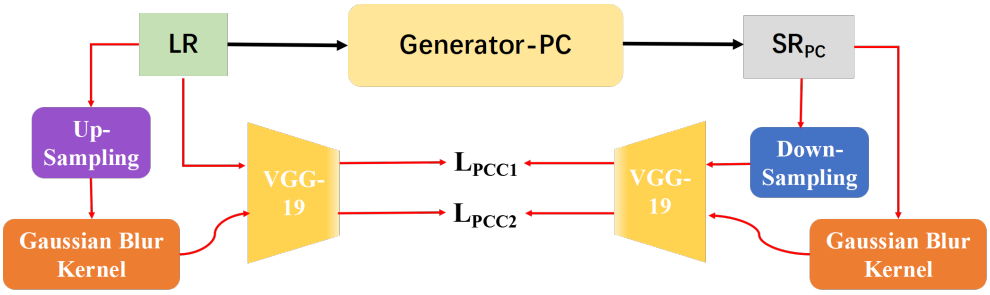

Figure 4. Overflow of the second stage with perceptual-content consistency.

In addition, even with noise and distortion in low-quality remote sensing images, low-level PC features should also be helpful for reconstructing desired structure details. So we proposed the perceptual down-scaling which is still calculated by using the L_1 norm.

$$L_{PCC_2} = \left\| \left(\varphi_{shallow}(D(I_{SR_{PC}})) - \varphi_{shallow}(I_L) \right) \right\|_1 \tag{13}$$

where $\varphi_{shallow}$ denotes shallow-level features "$Conv_{2,2}$" of VGG-19, which is utilized for perceptual representations of down-scaled $I_{SR_{PC}}$ and I_L. However, in calculating L_{PCC_2} for perceptual down-scaling similarity, different from L_{PCC_1}, the input of VGG-19 is the down-scaling $I_{SR_{PC}}$ and the original low-quality image I_L.

(3) Third stage: Perceptual-Distribution Consistency

Aiming at pursuing better visual effect, GAN [17] could reconstruct the photo-realistic details of remote sensing images with better perceptual performance due to the statistic gaming of the generator and discriminator. So the unpaired remote sensing ground-truths as real images I_r and the reconstructing results of the first two stages as fake ones I_f, similar to ESRGAN [18]. As shown in Figure 1, we employ the relativistic GAN to train the photo-realistic model to implement the perceptual-distribution consistency constraint. The relativistic discriminator intends to estimate the probability that I_r is more realistic than I_f. So L_{PDC} is composed of L_{Dis} and L_{Gen}. The corresponding adversarial loss is formulated as

$$L_{Dis} = -\mathbb{E}_{I_r}[\log(Dis_{Ra}(I_r, I_f))] - \mathbb{E}_{I_l}[\log(1 - Dis_{Ra}(I_f, I_r))] \tag{14}$$

The generator loss is calculated by

$$L_{Gen} = -\mathbb{E}_{I_r}[\log(1 - Dis_{Ra}(I_r, I_f))] - \mathbb{E}_{I_l}[\log(Dis_{Ra}(I_f, I_r))] \tag{15}$$

where Dis_{Ra} denotes the relativistic discriminator in ESRGAN [18] and loss functions are calculated by the cross-entropy. The framework of the discriminator is the the same as RankSRGAN [35]. In addition, we also employ the total variation loss L_{TV} as a regularization term to remove noises and checkerboard artifacts from images and keep the edges.

2.2. Dataset and Implementation Details

2.2.1. Dataset

We perform experiments to evaluate our method on three widely used benchmark datasets in remote sensing: WHU-RS19 [23], UC Merced dataset [36] and DOTA [37]. For WHU-RS19, UC Merced dataset, we randomly extract 70% original images for training as low-quality LR input, 15% images for validation, and the rest for testing. Especially, for the

unpaired super-resolution task, since the HR-sized inputs are needed in the third phase, 500 images from the DOTA dataset are selected as the unpaired HR references. To be fair, the test images are not used when training network as shown in Figure 5.

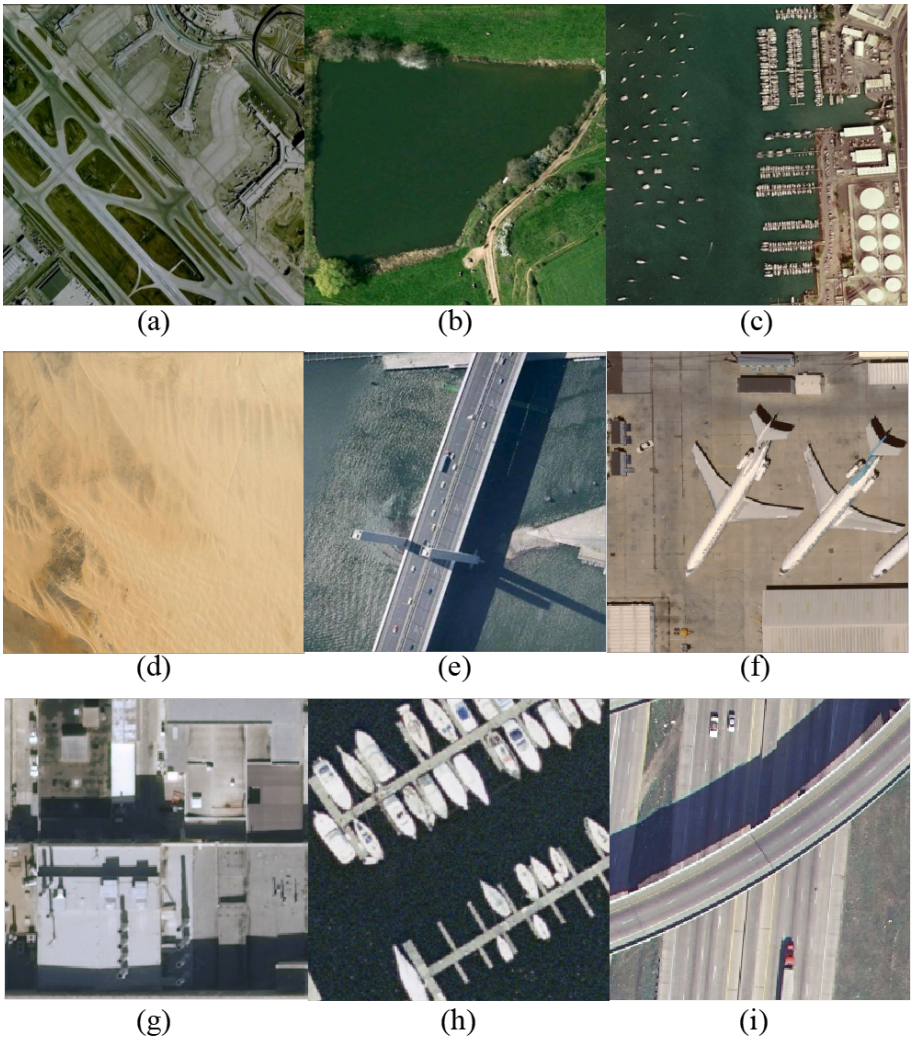

Figure 5. Part of test images chosen from WHU-RS19 and UC Merced test sets. (**a**) Airport_41. (**b**) Pond_43. (**c**) Port_45. (**d**) Desert_43. (**e**) Bridge_41. (**f**) airplane82. (**g**) buildings87. (**h**) harbor34. (**i**) overpass94.

2.2.2. Implementation Details and Metrics

As depicted in Section 3, in the MLFM structure, we stack EFEMs number as $n = 8$ to build a deep network with having 64 initial channels. In each EFEM, we stack sixteen residual blocks($m = 16$) which convolutional layers have 64 filters with the growth rate of 64. In the training process, we use the 74×74 RGB image patches which are randomly cropped from the low-quality training set, and then augment them with random horizontal flipping and rotation by 90°. Besides, the high-resolution high-quality patches with a size of 296×296 are randomly selected from the ground-truth. Furthermore, all the LR and HR images are pre-processed by subtracting the mean RGB value of the dataset for accelerating the ing phase. We train all of our models by using the Adam optimizer [38] with setting the momentum parameter $\beta_1 = 0.9$, $\beta_1 = 0.99$. The mini-batch size is set to 16 and the initial learning rate is 1×10^{-4} and halved for every 1×10^5 mini-batch updates. Each of the final models will get convergence after 2×10^5 mini-batch updates on PyTorch framework and the NVIDIA Titan RTX with 24 GB memory.

For evaluating the SR performance, we apply two common image quality assessment criteria for evaluating pixel-wise discrepancies: Peak Signal-to-Noise Ratio (PSNR) and Structural Similarity (SSIM) [39]. In addition, we also use a no-reference image quality assessment Natural Image Quality Evaluator (NIQE) [40] and a full-reference metric the Learned Perceptual Image Patch Similarity (LPIPS) [41] for scoring the degree of perceptual performance. Following the convention of super-resolution, only the luminance channel is selected for full-reference image quality assessment because the intensity of image is more sensitive to human-vision than the chroma.

3. Results

In this section, for assessing our method performance, we will give a detailed description of our experimental results. The proposed method is compared with the supervised methods and unsupervised methods on four common image quality assessments that we apply in Section 2.2.2.

3.1. Model Analysis

From Section 2.1.2, we train the final generator of our model is trained under the optimization of different stages with Content Consistency, Perceptual-Content Consistency and Perceptual-Distribution Consistency. In this part, we design the ablation study and analyze for investigating the effectiveness of each stage by training the ablated models.

In order to verify whether the feature maps extracted by the CC, PCC, and PDC have a dependency relationship, we show the visual results of training intermediate feature maps. As illustrated in Figure 6, we can detect that the feature maps extracted by the generator of different stages are similar. Therefore, our multi-stage aggregation network is reasonable because it can gradually reconstruct high-efficiency remote sensing images. To verify the necessity of using (muti-stage) architecture on unpaired dataset training, we remove CC, PCC and PDC from MSAN. As shown in Table 1 and Figure 7, specifically, those ablation study results demonstrate progressive consistency plays significant roles in qualitative performance:

Table 1. Ablation study on WHU-RS19 test dataset.

Method	Scale	PSNR↑	SSIM↑	NIQE↓	LPIPS↓
HR Ground Truth	-	-	-	4.541	-
Bicubic	×4	28.06	0.7233	8.005	0.4181
MSAN w/o PCC & PDC	×4	27.95	0.7067	8.958	0.3647
MSAN w/o PDC	×4	27.61	0.7285	8.219	0.3633
MSAN w/o CC	×4	26.83	0.7013	7.569	0.3677
MSAN	×4	27.43	0.7157	6.569	0.3511

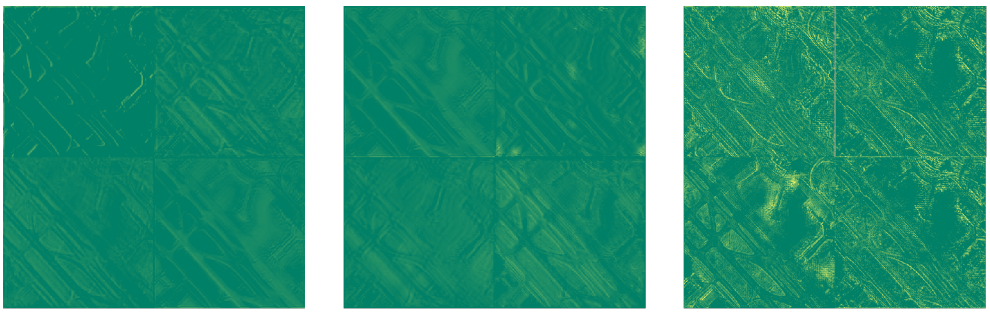

Figure 6. The feature maps of CC, PCC, PDC are visualized from left to right.

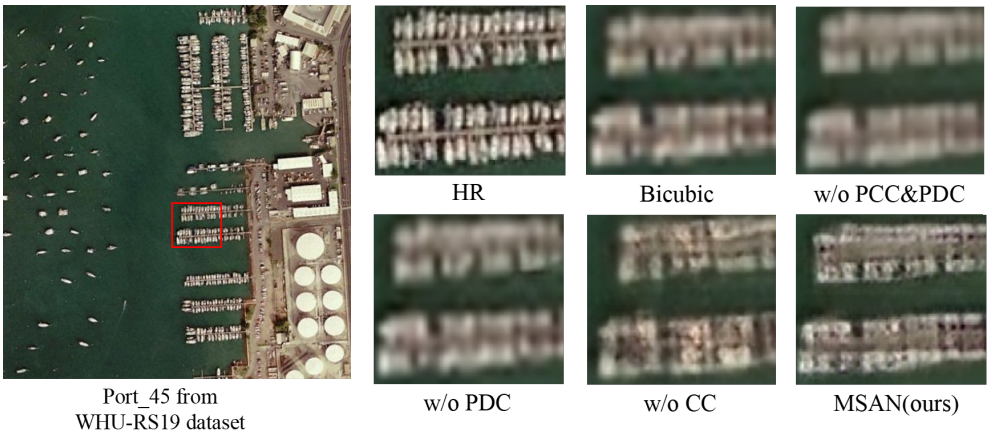

Figure 7. Visualization results of ablation study for ×4 upscaling on WHU-RS19 test set.

3.1.1. The Effect of Content Consistency

As depicted in Section 2.1.2(1), in the CC stage, the content-consistency loss L_{CC} is introduced for reconstructing the principal information from the low-resolution remote sensing inputs by down-scaling the super-resolved images. As Equation (11), this term aims at driving the degraded super-resolved results to be similar to the blurred low-quality inputs with the assumption of noiselessness. The principal content information of super-resolved results I_{SR_c} show high consistency to the low-resolution inputs as illustrated in Figure 7. That is why our model in the first stage has higher PSNR scores and lower perceptual metrics than other ablated model in Table 1.

3.1.2. The Effect of Perceptual-Content Consistency

As depicted in Section 2.1.2(2), by utilizing the pre-trained VGG-19 network to excavate the deep layer and shallow layer features, the effects of PCC that employs L_{PCC_1} and L_{PCC_2} should focus more on perceptual content information similarity between the feature level criteria of super-resolved results and low-resolution inputs. And as shown in Figure 4, L_{PCC_1} discriminates the deep layer features similarity with the assumptions of noiselessness using Gaussian blur $G(\cdot)$. L_{PCC_2} is applied on optimizing the distance of the shallow layer between the low-resolution inputs and down-sampling super-resolved results. Since the deep features extracted from the VGG-19 [34] depends on the local receptive fields strongly, the undesired noises would introduce some negative effects on the clear pixels to retain the perceptual-content consistency of I_{SR_C} and I_L. L_{PCC_2} mainly controls the similarity of principal information and explores perceptual to generate novel details as the perceptual features of noises have positive effects on information supplement of details. As shown in Figure 8, to evaluate the effect of different Gaussian kernels in computing the perceptual consistency, we conduct the experimental comparisons. The results suggest that the size of kernel (5×5) and the standard deviation value ($\sigma = 1.2$) are proper. Furthermore, as depicted in Figures 6 and 7, adding the PCC stage to the model focuses more on the perceptual content consistency of super-resolved results from the corresponding inputs. Due to remote sensing images containing the complex distribution of objects, the PCC stage mainly avoids lack of informative details, over-fitting on edges and artifacts.

3.1.3. The Effect of Perceptual-Distribution Consistency

As depicted in Section 2.1.2(3), we exploit the adversarial learning to train the perceptual model for photo-realistic super-resolution. Especially for experiments, as shown in Table 1, comparing the model MSAN w/o PC and MSAN (NIQE and LPIPS), it fully demonstrated that perceptual-distribution consistency is beneficial to improve the visual effect. From Figure 7, we can observe that the super-resolved results of the MSAN without

PDC cannot generate photo-realistic textures and details. It means PDC is necessary for our architecture.

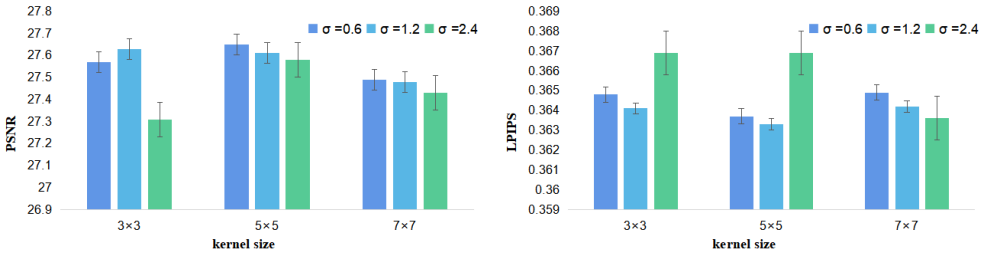

Figure 8. Experimental comparisons of differnet Gaussian kernels for ×4 upscaling on WHU-RS19 test set.

3.2. Comparisons with State-of-the-Art Methods

For a comprehensive comparison, we compare our method with several paired methods including SRCNN [9], EDSR [15], ESRGAN [18], RCAN [19], IMDN [42]and HAN [43], and unpaired methods including SelfExSR [24], ZSSR [26], CinCGAN [25] and DRN [44]. These methods are all evaluated on the UC Merced test set and the WHU-RS19 test set with bicubic degradation and complex degradations (bicubic + Gaussian blur) at the scaling factor of ×4. Bicubic degradation is the most widely used assumption in the paired SR task, although it cannot effectively fit the original remote sensing image degradation model. Therefore, we conducted experiments on bicubic degradation as a baseline and add blur degradation to analyze the robustness of MSAN against complex degradation.

3.2.1. Result on Bicubic Degradation

In this part, two remote sensing test sets are down-sampled with bicubic degradation. Tables 2 and 3 show quantitative comparisons for ×4 SR and present the average assessment results.

Table 2 presents the quantitative results of evaluation between our method and other unpaired methods. Consider the UC Merced test set as an example. It can be seen that our method achieves the best performance in all cases. Specifically, the PSNR gains which MSAN exceeds the second-best model reach 0.48 dB. The LPIPS of our method is 0.0062 lower than that obtained with the suboptimal method. The proposed MSAN shows significant advantages of NIQE. The NIQE and LPIPS of the MSAN are both optimal compared with those of other unpaired methods on the WHU-RS19 test set. The NIQE and LPIPS of our method are 0.047 and 0.0036 lower than that obtained with the suboptimal method respectively. According to Table 2, in the remote sensing dataset, the unpaired method we proposed is superior to all the other unpaired methods on most of the metrics. It is verified that MSAN can reconstruct more subjective perception remote sensing images.

Table 3 presents the quantitative results of evaluation between our method and paired methods. As for the ×4 SR task on UC Merced test set, the cases of our method all achieve the best performance. The proposed MSAN performs the best in terms of LPIPS and NIQE and keeps the presentable PSNR values. Taking the WHU-RS19 test set as an example. For bicubic degradation, experimental results demonstrate that the proposed method outperforms part of paired SR methods. Our method achieves the best LPIPS performance with those of other methods. The LPIPS is 0.009 lower than that obtained with the suboptimal IMDN method. HAN keeps the presentable PSNR and SSIM values. Although ESRGAN achieves the lowest NIQE, the content of its super-resolution results is often inconsistent by undesirable artifacts so that the PSNR of the ESRGAN is the lowest as shown in Figure 9.

As shown in Figure 9, we present visual comparisons on bicubic degradation datasets with the Bicubic technique and other state-of-the-art methods, which are among the methods listed in Table 2 and 3. Here, we use SSIM and non-reference image quality metrics NIQE to assess these super-resolved results. From Figure 9a, we observe that most of the

compared unpaired methods suffer from blurred artifacts so that it is difficult to reconstruct the correct content and structure edge. The SelfExSR and DRN reconstruct SR image contains wrong details that can't match the information in the groud-truth images. In contrast, our MSAN can slightly alleviate this phenomenon and produce more accurate results with more precise details. In Figure 9b, we present visual comparisons of some paired methods reconstruction results. For image "Desert_43" from the WHU-RS19 dataset, our MSAN can synthesize realistic textures while retaining a delicate content compared to other methods.

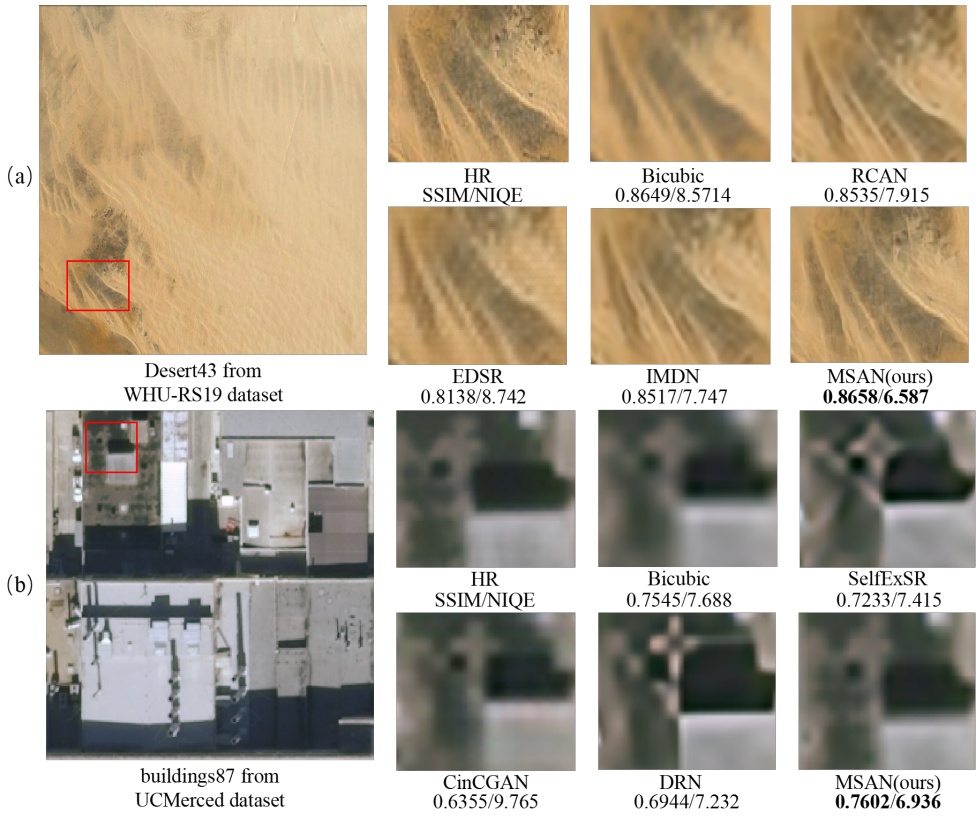

Figure 9. Visual comparisons of MSAN with other methods on bicubic degradation for ×4 upscaling. (**a**) Super-resolved results from WhU-RS19 test set. (**b**) Super-resolved results from UCMerced test set.

Table 2. Average resultes of unpaired SR methods on bicubic degradation. The best results were highlighted with bold black.

Dataset	Method	Scale	PSNR	SSIM	NIQE	LPIPS
UC Merced	SelfExSR [24]	×4	21.40	0.5698	10.01	0.6882
	ZSSR [26]	×4	22.67	0.6156	9.690	0.6850
	CinCGAN [25]	×4	23.46	0.6355	9.756	0.4259
	DRN [44]	×4	22.05	0.5477	12.69	0.6632
	MSAN (ours)	×4	**23.94**	**0.6451**	**7.207**	**0.4197**
WHU-RS19	SelfExSR [24]	×4	28.12	0.7398	6.625	0.3660
	ZSSR [26]	×4	25.29	0.6882	7.667	0.4009
	CinCGAN [25]	×4	26.47	0.7073	6.516	0.3740
	DRN [44]	×4	**28.85**	**0.7693**	7.298	0.3547
	MSAN (ours)	×4	27.43	0.7157	**6.403**	**0.3511**

Table 3. Average resultes of paired SR methods on bicubic degradation. The best results were highlighted with bold black.

Dataset	Method	Scale	PSNR	SSIM	NIQE	LPIPS
UC Merced	SRCNN [9]	×4	22.25	0.5978	10.51	0.4435
	EDSR [15]	×4	23.71	0.6412	9.219	0.6822
	ESRGAN [18]	×4	19.65	0.5192	18.90	0.6642
	RCAN [19]	×4	19.71	0.5207	14.08	0.6632
	IMDN [42]	×4	20.39	0.5484	11.73	0.6704
	HAN [43]	×4	20.04	0.5419	12.67	0.6589
	MSAN (ours)	×4	**23.94**	**0.6451**	**7.207**	**0.4197**
WHU-RS19	SRCNN [9]	×4	28.15	0.7414	6.891	0.3526
	EDSR [15]	×4	27.83	0.7238	7.533	0.3890
	ESRGAN [18]	×4	25.62	0.6268	**4.338**	0.3724
	RCAN [19]	×4	28.74	0.7670	7.006	0.3535
	IMDN [42]	×4	28.65	0.7612	7.212	0.3520
	HAN [43]	×4	**28.89**	**0.7705**	7.247	0.3554
	MSAN (ours)	×4	27.43	0.7157	6.403	**0.3511**

3.2.2. Result on Complex Degradation

We further apply our MSAN to solve the complex degradation image super-resolution to analyze the robustness. In order to construct complex degradation, the isotropic Gaussian blur is added to low-resolution images. Quantitative comparisons for $\times 4$ SR are shown in Tables 4 and 5 and visual comparisons are shown in Figure 10.

In Table 4, compared with other unpaired SR methods, our method achieves the best quantitative results of the evaluation. Making the UC Merced test set as an example, It can be seen that our method achieves the best performance on PSNR, SSIM and NIQE. CinCGAN and DRN also render competitive results. Specifically, The LPIPS of CinCGAN is 0.4487 lower than other unpaired methods. The PSNR and SSIM which MSAN exceeds the second-best model—DRN reach 0.02 dB and 0.0101 respectively. The proposed MSAN still shows significant advantages of NIQE. In the WHU-RS19 test set, the SSIM and NIQE of the MSAN are both optimal compared with those of other unpaired methods. The PSNR of our method is 0.04 dB lower than that obtained with the best method.

Table 4. Average resultes of unpaired SR methods on degradation on complex degradation. The best results were highlighted with bold black.

Dataset	Method	Scale	PSNR	SSIM	NIQE	LPIPS
UC Merced	SelfExSR [24]	×4	23.98	0.6109	9.417	0.7271
	ZSSR [26]	×4	23.79	0.6059	9.111	0.7236
	CinCGAN [25]	×4	23.56	0.6156	9.369	**0.4487**
	DRN [44]	×4	24.11	0.6214	10.84	0.7182
	MSAN (ours)	×4	**24.13**	**0.6315**	**7.855**	0.7214
WHU-RS19	SelfExSR [24]	×4	26.42	0.6916	6.614	0.3993
	ZSSR [26]	×4	**26.65**	0.7052	7.229	**0.3896**
	CinCGAN [25]	×4	25.68	0.6898	6.454	0.4011
	DRN [44]	×4	26.58	0.6835	7.548	0.4036
	MSAN (ours)	×4	26.61	**0.7183**	**5.112**	0.3936

Table 5. Average resultes of paired SR methods on complex degradation. The best results were highlighted with bold black.

Dataset	Method	Scale	PSNR	SSIM	NIQE	LPIPS
UC Merced	SRCNN [9]	×4	23.18	0.6403	9.305	**0.4207**
	EDSR [15]	×4	24.10	0.6301	8.680	0.7089
	ESRGAN [18]	×4	21.11	0.4683	7.860	0.6340
	RCAN [19]	×4	23.96	0.6481	15.02	0.6810
	IMDN [42]	×4	22.97	0.5365	11.12	0.7736
	HAN [43]	×4	24.07	**0.6511**	10.64	0.6152
	MSAN (ours)	×4	**24.13**	0.6315	**7.855**	0.7214
WHU-RS19	SRCNN [9]	×4	25.37	0.6533	6.018	0.4354
	EDSR [15]	×4	26.53	0.7011	7.366	0.3911
	ESRGAN [18]	×4	25.08	0.5910	**4.583**	0.4402
	RCAN [19]	×4	26.59	0.7127	6.894	**0.3876**
	IMDN [42]	×4	26.59	0.7085	7.007	0.3901
	HAN [43]	×4	26.55	0.6845	7.544	0.4171
	MSAN (ours)	×4	**26.61**	**0.7183**	5.112	0.3936

Figure 10. Visual comparisons of MSAN with other SR mathods on complex degradation for ×4 upscaling. (**a**,**b**) Super-resolved results of different images from WHU-RS19 test set.

As shown in Table 5, Several evaluation indexes render competitive results in contrast with paired SR methods. Compared to ESRGAN, the proposed MSAN achieves the best in terms of PSNR and keeps the presentable LPIPS and NIQE values. Although SRCNN achieves the lowest LPIPS, NIQE of it is the highest compared to other methods. To show that the MSAN has an excellent reconstruction effect on blur low-resolution images, we present the visual effect in Figure 10, where the images are from WHU-RS19. From Figure 10, ESRGAN generates many distorted textures on complex degradation test sets; in addition, ZSSR and SelfExSR suffer from the jagged effect on the edges of details. Super-

resolved results of our method reconstruct more content details than other methods. This firmly reflects the strong SR capacity of our models in dealing with complex degradation.

4. Discussion

4.1. Method of Application

In Section 3, all experiments are based on the down-sampling of the original images with a known degradation model to obtain low-quality low-resolution images of the test set. However, in the real world, the information transportation and compression procedure are generally unknown. Therefore, reconstructing original low-quality images of different datasets to further discuss the proposed method's SR reconstruction effect is necessary.

To this end, the sample without degradation operation is chosen from the UC Merced dataset with a spatial resolution of 0.3 m per pixel. The methods based on the Bicubic, ZSSR, and MSAN are used to perform SR reconstruction. As shown in Figure 11a, it presents the results of SR reconstruction with an up-sampling factor $\times 4$. Due to the lack of corresponding ground-truth images, the reconstruction performance can be evaluated by comparing the visual effect reconstructed by different methods. In addition, we chose three original images from the DOTA and GaoFen-1 with a high spatial resolution as shown in Figure 11b–d.

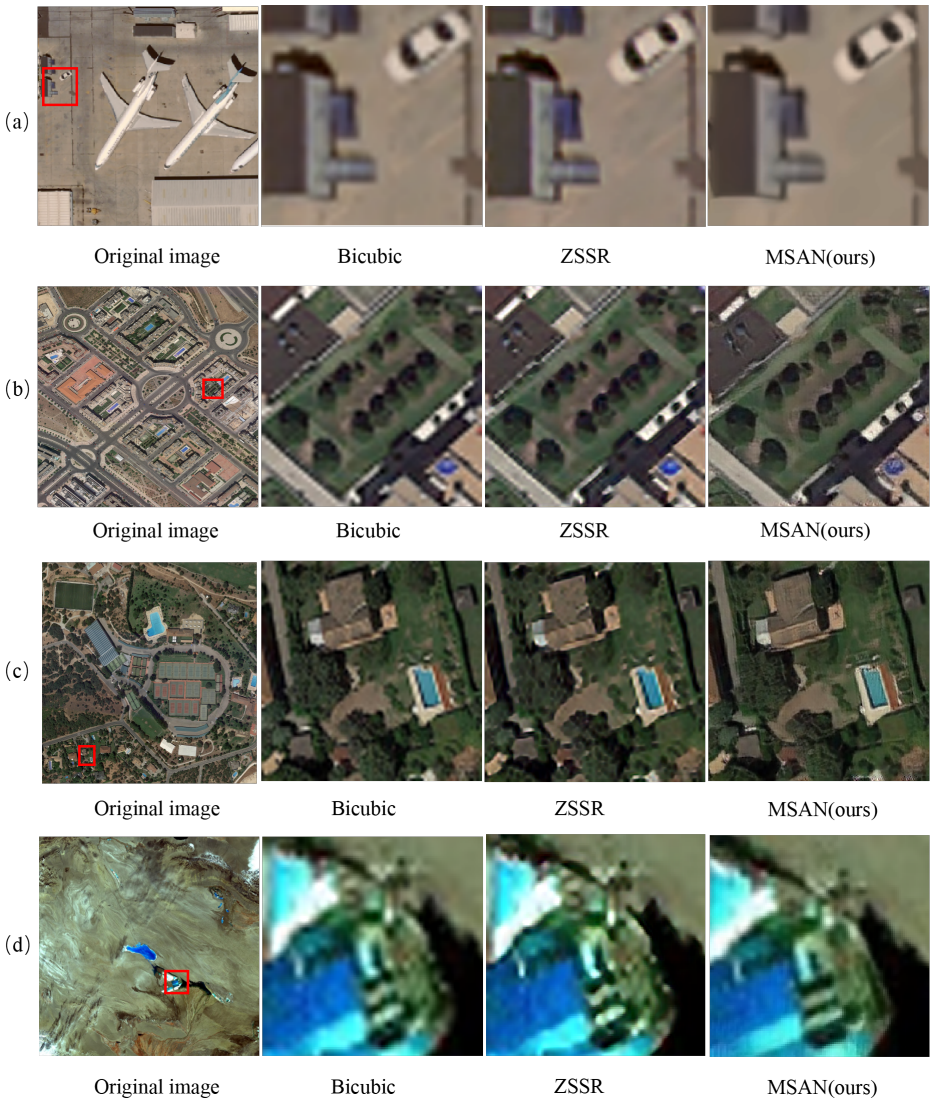

Figure 11. Visual comparisons of the Bicubic model, ZSSR, and MSAN when applied to original remote sensing images in each group of pictures. (**a**) Super-resolved results from UC-Merced test set. (**b**,**c**) Super-resolved results from DOTA dataset. (**d**) Super-resolved results from GaoFen-1.

Because of the quality of different original images, the visual effect of the image reconstructed by our method on the first two images is not as good as that on the last two images. Nevertheless, the proposed method obtains better performance with more details and textures and more precise contour stripes than the two other methods. These experimental results further verify that our MSAN produces promising results when tackling real-world remote sensing data with complicated patterns.

4.2. Limitation

By quantitatively analyzing the experimental results on various datasets, we find that our MSAN achieves the best performance on the WHU-RS19 dataset but only reaches the second-best on the UC Merced dataset, whose high-resolution counterparts have lower quality and resolution than the WHU-RS19 dataset. In a sense, it indicates that our MSAN shows less effectiveness and generalization on reconstructing small-sized images. For a reason, in the content reconstruction stage, the content consistency constraint is applied to optimize the model by calculating the pixel-wise discrepancy. And in the UC Merced dataset, the images with the size of 256×256 contain fewer contents for reconstruction. Due to the progressive reconstruction, only if the super-resolved results of the first stage preserve the principal information of the image, the subsequent stages would refine the perceptual textures well. Therefore, in the future, it is expected to improve the capability of the model and optimization for content consistency in the first stage.

5. Conclusions

In this paper, we propose an efficient unpaired super-resolution method with multi-stage aggregation network(MSAN) to super-resolved real remote sensing images. Specifically, we utilize content consistency, perceptual-content consistency and perceptual-distribution consistency to exert a stage-by-stage training mode, which helps reconstruct accurate content, avoid structure details loss and generate desired photo-realistic details. We introduce different consistency losses for different phases to optimize content and distribution similarity between the super-resolved results and low-quality inputs. The generator of MSAN mainly includes EFEMs to improve the use of original low-quality image feature information. The experiments of method analysis suggest the effectiveness of progressive reconstruction strategy and every phase is dependent. Extensive experimental results demonstrate that our method achieves satisfactory performance in four metrics compared with current unpaired SR methods and renders a competitive visual effect in contrast with paired SR methods. Besides, we conduct our experiments on original remote sensing images and obtains excellent performance, proving the robustness and practicality in real-world applications.

Author Contributions: Conceptualization, L.Z.; methodology, L.Z., W.L. and Y.H.; software, L.Z.; validation, L.Z., X.S. and Y.H.; formal analysis, L.Z. and W.L; investigation, L.Z.; resources, L.Z. and H.Z.; data curation, L.Z.; writing—original draft preparation, L.Z.; writing—review and editing, L.Z. and Y.H.; visualization, L.Z.; supervision, W.L.; project administration, W.L.; funding acquisition, W.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported in part by the National Natural Science Foundation of China (Grant No. 61871311), the Key Industrial Innovation Chain Project in Industrial Domain of Shaanxi Province (Grant No. 2020ZDLGY05-01), Aeronautical Science Foundation of China (Grant No. 2020Z071081004).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data presented in this study are available on request from the corresponding author.

Conflicts of Interest: All authors have reviewed the manuscript and approved submission to this journal. The authors declare that there is no conflict of interest regarding the publication of this article and no self-citations included in the manuscript.

References

1. Shirkolaei, M.M. High efficiency X-band series-fed microstrip array antenna. *Prog. Electromagn. Res. C* **2020**, *105*, 35–45. [CrossRef]
2. Alibakhshikenari, M.; Babaeian, F.; Virdee, B.S.; Aïssa, S.; Azpilicueta, L.; See, C.H.; Althuwayb, A.A.; Huynen, I.; Abd-Alhameed, R.A.; Falcone, F.; et al. A Comprehensive Survey on "Various Decoupling Mechanisms With Focus on Metamaterial and Metasurface Principles Applicable to SAR and MIMO Antenna Systems". *IEEE Access* **2020**, *8*, 192965–193004. [CrossRef]
3. Qin, M.; Hu, L.; Du, Z.; Gao, Y.; Qin, L.; Zhang, F.; Liu, R. Achieving Higher Resolution Lake Area from Remote Sensing Images through an Unsupervised Deep Learning Super-Resolution Method. *Remote. Sens.* **2020**, *12*, 1937. [CrossRef]
4. Allebach, J.; Wong, P.W. Edge-directed interpolation. In Proceedings of the IEEE International Conference on Image Processing, Lausanne, Switzerland, 16–19 September 1996; pp. 707–710.
5. Dong, W.; Zhang, L.; Shi, G.; Wu, X. Nonlocal back-projection for adaptive image enlargement. In Proceeding of the IEEE International Conference on Image Processing , Cairo, Egypt, 7–10 November 2009; pp. 349–352.
6. Yang, J.; Wright, J.; Huang, T.; Ma, Y. Image super-resolution as sparse representation of raw image patches. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Anchorage, AK, USA, 23–28 June 2008; pp. 1–8.
7. Gu, S.; Zuo, W.; Xie, Q.; Meng, D.; Feng, X.; Zhang, L. Convolutional sparse coding for image super-resolution. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 7–13 December 2015; pp. 1823–1831.
8. Peng, C.; Gao, X.; Wang, N.; Li, J. Graphical representation for heterogeneous face recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **2016**, *39*, 301–312. [CrossRef] [PubMed]
9. 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]
10. Kim, J.; Lee, J.K.; Lee, K.M. Accurate image super-resolution using very deep convolutional networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 1646–1654.
11. 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.
12. 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, 27–30 June 2016; pp. 770–778.
13. Zhang, Y.; Tian, Y.; Kong, Y.; Zhong, B.; Fu, Y. Residual dense network for image super-resolution. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 2472–2481.
14. 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.
15. 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.
16. 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.
17. Goodfellow, I.J.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; Bengio, Y. Generative adversarial networks. A. Courville, and Y. Bengio, Generative adversarial networks. In Proceedings of the International Conference on Neural Information Processing,Kyoto, Japan, 16–21 October 2016; pp. 2672–2680.
18. Wang, X.; Yu, K.; Wu, S.; Gu, J.; Liu, Y.; Dong, C.; Qiao, Y.; Change Loy, C. Esrgan: Enhanced super-resolution generative adversarial networks. In Proceedings of the European Conference on Computer Vision Workshops, Munich, Germany, 8–14 September 2018; pp. 63–79.
19. 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.
20. Huang, Y.; Li, J.; Gao, X.; Hu, Y.; Lu, W. Interpretable Detail-Fidelity Attention Network for Single Image Super-Resolution. *IEEE Trans. Image Process.* **2021**, *30*, 2325–2339. [CrossRef] [PubMed]
21. Efrat, N.; Glasner, D.; Apartsin, A.; Nadler, B.; Levin, A. Accurate blur models vs. image priors in single image super-resolution. In Proceedings of the IEEE International Conference on Computer Vision, Sydney, NSW, Australia, 1–8 December 2013; pp. 2832–2839.
22. Yang, C.Y.; Ma, C.; Yang, M.H. Single-image super-resolution: A benchmark. In Proceedings of the European Conference on Computer Vision, Zurich, Switzerland, 6–12 September 2014; pp. 372–386.
23. Dai, D.; Yang, W. Satellite image classification via two-layer sparse coding with biased image representation. *IEEE Geosci. Remote Sens. Lett.* **2010**, *8*, 173–176. [CrossRef]
24. Freedman, G.; Fattal, R. Image and video upscaling from local self-examples. *ACM Trans. Graph. (TOG)* **2011**, *30*, 474-484. [CrossRef]
25. Yuan, Y.; Liu, S.; Zhang, J.; Zhang, Y.; Dong, C.; Lin, L. Unsupervised image super-resolution using cycle-in-cycle generative adversarial networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, Salt Lake City, UT, USA, 18–22 June 2018; pp. 701–710.

26. 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, 18–23 June 2018; pp. 3118–3126.
27. 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, 18–23 June 2018; pp. 9446–9454.
28. Zhang, N.; Wang, Y.; Zhang, X.; Xu, D.; Wang, X. An unsupervised remote sensing single-image super-resolution method based on generative adversarial network. *IEEE Access* **2020**, *8*, 29027–29039. [CrossRef]
29. Wang, P.; Zhang, H.; Zhou, F.; Jiang, Z. Unsupervised remote sensing image super-resolution using cycle CNN. In Proceedings of the IGARSS 2019 IEEE International Geoscience and Remote Sensing Symposium, Yokohama, Japan, 28 July–2 August 2019; pp. 3117–3120.
30. Zhang, N.; Wang, Y.; Zhang, X.; Xu, D.; Wang, X.; Ben, G.; Zhao, Z.; Li, Z. A Multi-Degradation Aided Method for Unsupervised Remote Sensing Image Super Resolution with Convolution Neural Networks. *IEEE Trans. Geosci. Remote Sens.* **2020**, 1–14. [CrossRef]
31. Huang, Y.; Sun, X.; Lu, W.; Li, J.; Gao, X. Un-Paired Real World Super-Resolution with Degradation Consistency. In Proceedings of the IEEE International Conference on Computer Vision Workshops, Seoul, Korea, 27–28 October 2019; pp. 3458–3466.
32. 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, Honolulu, HI, USA, 21–26 July 2017; pp. 4700–4708.
33. Shi, W.; Caballero, J.; Huszár, F.; Totz, J.; Aitken, A.P.; Bishop, R.; Rueckert, D.; Wang, Z. Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 1874–1883.
34. Simonyan, K.; Zisserman, A. Very deep convolutional networks for large scale image recognition. In Proccedings of the 3rd International Conference on Learning Representations, San Diego, CA, USA, 7–9 May 2015
35. Zhang, W.; Liu, Y.; Dong, C.; Qiao, Y. RankSRGAN: Generative adversarial networks with ranker for image Super-resolution. In Proceedings of the IEEE International Conference on Computer Vision, Seoul, Korea, 27 October–2 November 2019.
36. Yang, Y.; Newsam, S. Bag-Of-Visual-Words and Spatial Extensions for Land-Use Classification. In Proceedings of the 18th SIGSPATIAL International Conference on Advances in Geographic Information Systems, San Jose, CA, USA, 2–5 November 2010; pp. 270–279. [CrossRef]
37. Xia, G.S.; Bai, X.; Ding, J.; Zhu, Z.; Belongie, S.; Luo, J.; Datcu, M.; Pelillo, M.; Zhang, L. DOTA: A large-scale dataset for object detection in aerial images. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 3974–3983.
38. Kingma, D.P.; Ba, J.L. Adam: A method for stochastic optimization. In Proceedings of the 3rd International Conference on Learning Representations, San Diego, CA, USA, 7–9 May 2015.
39. 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]
40. Mittal, A.; Soundararajan, R.; Bovik, A.C. Making a "completely blind" image quality analyzer. *IEEE Signal. Process. Lett.* **2012**, *20*, 209–212. [CrossRef]
41. Zhang, R.; Isola, P.; Efros, A.A.; Shechtman, E.; Wang, O. The unreasonable effectiveness of deep features as a perceptual metric. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 586–595.
42. Hui, Z.; Gao, X.; Yang, Y.; Wang, X. Lightweight image super-resolution with information multi-distillation network. In Proceedings of the MM'19: 27th ACM International Conference on Multimedia, Nice, France, 21–25 October 2019; pp. 2024–2032.
43. Niu, B.; Wen, W.; Ren, W.; Zhang, X.; Yang, L.; Wang, S.; Zhang, K.; Cao, X.; Shen, H. Single image super-resolution via a holistic attention network. In Proceedings of the European Conference on Computer Vision, Glasgow, UK, 23–28 August 2020; pp. 191–207.
44. Guo, Y.; Chen, J.; Wang, J.; Chen, Q.; Cao, J.; Deng, Z.; Xu, Y.; Tan, M. Closed-loop matters: Dual regression networks for single image super-resolution. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 5407–5416.