SCIENCE CHINA
Information Sciences

• Supplementary File •

Unpaired remote sensing image super resolution with content-preserving weak supervision neural network

Jie WU[†1], Runmin CONG[†2], Leyuan FANG[*1], Chunle GUO[3], Bob ZHANG[4] & Pedram GHAMISI[5,6]

[1]*College of Electrical and Information Engineering, Hunan University, Changsha* 410082*, China;*
[2]*Institute of Information Science, Beijing Jiaotong University, Beijing* 100044*, China;*
[3]*College of Computer Science, Nankai University, Tianjin* 300071*, China;*
[4]*Department of Computer and Information Science, University of Macau, Macau 999078, China;*
[5]*Helmholtz-Zentrum Dresden-Rossendorf, Helmholtz Institute Freiberg for Resource Technology, Freiberg* 09599*, Germany;*
[6]*Institute of Advanced Research in Artificial Intelligence (IARAI), Landstraer Hauptstrae 5, Vienna* 1030*, Austria*

Appendix A Related work

Remote sensing image SR aims to recover detailed HR images from LR images. Generally, the HR images I^{HR} are degraded to LR images I^{LR} by a degradation model:

$$I^{LR} = (I^{HR} \otimes k) \downarrow_s + n, \tag{A1}$$

where k is a blur kernel, and $\otimes$ is the convolution operation, and $\downarrow_s$ is downsampling operation with a scale factor of s, and n is the additional noise. For remote sensing images, the degradation model is unknown so that the SR is an ill-posed problem with infinite solutions. The difficulty of SR is to limit the solution space to obtain better SR results.

Appendix A.1 CNN-based remote sensing image SR

With the rapid development of deep learning methods, CNN-based remote sensing image SR methods [1] have achieved good performance. These methods aim to learn the mapping function $F(\cdot)$ from the paired HR and LR data sets:

$$I^{SR} = F(I^{LR}), \tag{A2}$$

where I^{SR} represents a reconstructed SR image and I^{LR} is downsampled from HR images I^{HR} by bicubic interpolation. The loss function of CNN-based SR generally uses the mean absolute error (MAE) to measure the difference between I^{SR} and I^{HR}:

$$L_{MAE}(F, I^{HR}, I^{LR}) = \left\| F(I^{LR}) - I^{HR} \right\|_1, \tag{A3}$$

where L_{MAE} represents the MAE loss function to optimize the CNN network parameters.

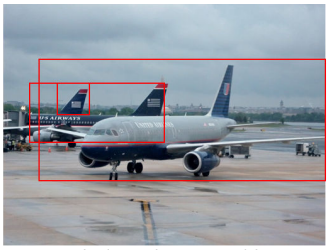
Recently, CNN-based methods have been applied for the remote sensing images SR. In [2], Lei et al. proposed a local-global-combined network (LGCNet) to learn the multi-level representations of remote sensing images including both local details and global environmental priors. A deep distillation recursive network (DDRN) [3] was introduced for video satellite image SR with ultra-dense residual blocks and a recursive strategy to mitigate memory consumption. Lu et al. [4] presented a multi-scale residual neural network (MRNN) by adopting the multi-scale nature of satellite images to reconstruct high-frequency information. Pan et al. [5] proposed a residual dense back-projection network (RDBPN) to enhance the resolution of RGB remote sensing images with dense back-projection blocks, which utilized residual learning in both global and local manners. In [6], Zhang et. al. proposed a mixed high-order attention network improved by a high-order attention module to restore the missing details. Driven by the dense connections [7], Dong et. al developed a dense-sampling SR network [8] named DSSR to achieve the large-scale SR reconstruction by using the dense-sampling mechanism and wide feature attention block. However, these CNN-based methods need paired data sets to train the SR network. The typical way for synthesizing the paired data is the utilization of the bicubic model that is quite different from the real-world remote sensing degradation model. Therefore, these CNN-based methods achieve poor performance when applied to real LR remote sensing images.

Appendix A.2 GAN-based remote sensing image SR

Recently, benefiting from the adversarial learning strategy, the GAN-based SR methods have achieved impressive results with more detailed information and visual effect in natural image SR [9] [10] [11] and remote sensing image SR. The GAN-based methods are typically composed of a generator and a discriminator. For SR tasks, the generator is used to reconstruct SR images and the discriminator aims to distinguish SR images from HR images. The loss function of GAN-based SR methods is as follows:

$$L_{adv}(G, D, I^{HR}, I^{LR}) = \log(D(I^{HR})) + \log(1 - D(G(I^{LR}))), \tag{A4}$$

* Corresponding author (email: leyuan_fang@hnu.edu.cn)
† Authors A and B have the same contribution to this work.



(a) Airplanes in a natural image

(b) Airplanes in a remote sensing image

(c) An SR reconstructed remote sensing image without content-preserving constraints

Figure A1 Comparison of airplanes in a natural image (a) and a remote sensing image (b). Image (c) is the SR reconstructed remote sensing image for (b) without content-preserving constraints. Airplanes are marked by red rectangles.

where G is the generator to learn the SR mapping function, and D is the discriminator. The min-max two-player game between the generator and the discriminator makes the generated data be indistinguishable from the real ones.

For remote sensing SR, most of the researchers improve the performance from two aspects: 1. the structure of the generator; 2. the judgment ability of the discriminator. Specifically, Kui et al. [12] proposed a GAN-based edge-enhancement network (EEGAN), where the generator was improved by an ultra-dense subnetwork and an edge-enhancement subnetwork to increase the robustness of satellite image SR reconstruction. In [13], Lei et al. proposed a coupled-discriminated GANs (CDGAN), in which the discriminator was specifically designed to take SR and HR images as inputs to improve the discrimination ability of the discriminator. Though GAN-based methods reconstruct SR results with fine texture and details, they still need paired data sets to training, which limits the generalization ability of the model. In addition, these methods are more sensitive to noise which results in the compensated high-frequency details (e.g., image edges) of SR results may be inconsistent with the ground truth images.

Appendix A.3 Image domain translation

Image domain translation aims to transform an input image (source domain) into a specific target image (target domain), e.g., mapping aerial images to map images [14]. Inspired by the GANs, the image domain translation methods have gained fruitful progress. In [15], the CycleGAN proposed by Zhu et al. can learn the translation between two different domains X and Y . The CycleGAN has two generators, $G_1 : X \to Y$ and $G_2 : Y \to X$ accompanied by two discriminators D_X and D_Y. In addition, the CycleGAN adopts the cycle consistency loss to reduce the space of possible mapping functions. For each image $x \in X$ and $y \in Y$, the cycle consistency loss can be expressed as:

$$L_{cyc}(G_1, G_2, x, y) = \|G_2(G_1(x)) - x\|_1 + \|G_1(G_2(y)) - y\|_1, \tag{A5}$$

where L_{cyc} describes the cycle consistency loss. The identity loss L_{idt} is applied to preserve color composition between input and output images and avoid the color variation issue, which is expressed as:

$$L_{idt}(G_1, G_2) = \left\|I^{LR}_{real,G_1} - I^{LR}_{real}\right\|_1 + \left\|I^{LR}_{bic,G_2} - I^{LR}_{bic}\right\|_1. \tag{A6}$$

Appendix B Training loss function

The full loss functions of image domain translation are shown as follows:

$$\begin{aligned} L_{pseudo-LR} = {} & L_{adv}(G_1, D_{real}, I^{LR}_{real}, I^{LR}_{bic}) + L_{adv}(G_2, D_{bic}, I^{LR}_{bic}, I^{LR}_{real}) + \\ & \lambda_1 L_{cyc}(G_1, G_2, I^{LR}_{bic}, I^{LR}_{real}) + \lambda_2 L_{idt}(G_1, G_2) + \lambda_3 L_{per}(G_1, G_2), \end{aligned} \tag{B1}$$

where λ_1, λ_2, λ_3 are hyperparameters to weight the contributions of each loss function. L_{adv} is the adversarial loss function described in Eq. (A4). L_{cyc} is the cycle consistency loss described in Eq. (A5). The identity loss is estimated by Eq. (A6). The perceptual loss is described as:

$$L_{per}(G_1, G_2) = \left\| \phi_{m,n}(I^{LR}_{pseudo}) - \phi_{m,n}(I^{LR}_{bic}) \right\|_1 + \left\| \phi_{m,n}(I^{LR}_{real,G_2}) - \phi_{m,n}(I^{LR}_{real}) \right\|_1, \tag{B2}$$

where $\phi_{m,n}$ denotes the feature map extracted by the n-th convolutional layer before m-th max pooling layer of the VGG19 extractor.

The full loss function of the SR reconstruction is composed of the MAE loss L_{MAE}, the edge retention loss L_{edge}, the degradation consistency loss L_{de}, and the adversarial loss L_{adv}:

$$L_{SR} = \omega_1 \cdot L_{MAE}(G_{SR}, I^{HR}_{real}, I^{LR}_{pseudo}) + \omega_2 \cdot L_{edge}(G_{edge}, D_{edge}) + \omega_3 \cdot L_{de}(G_{de}, D_{de}) + \omega_4 \cdot L_{adv}(G_{SR}, D_{hr}, I^{HR}_{real}, I^{LR}_{real}), \tag{B3}$$

where ω_1, ω_2, ω_3 , ω_4 are weights of each loss. L_{MAE} and L_{adv} are estimated by Eq. (A3) and Eq. (A4), respectively. L_{edge} is defined as:

$$L_{edge}(G_{edge}) = -[I^{LR}_{real,BIC\circ G_{edge}} \log(I^{SR}_{real,G_{edge}}) + (1 - I^{LR}_{real,BIC\circ G_{edge}}) \log(1 - I^{SR}_{real,G_{edge}})], \tag{B4}$$

where G_{edge} is the DexiNed edge detection network, and BIC represents the bicubic interpolation to upscale I^{LR}_{real} to the size of I^{SR}_{real}. $I^{LR}_{real,BIC\circ G_{edge}}$ represents the edge information extracted by the edge detection network from the bicubic upsampled I^{LR}_{real}. $I^{SR}_{real,G_{edge}}$ stands for the edge information of I^{SR}_{real}. L_{de} is described as:

$$L_{de}(G_{de}, D_{de}) = \log(D_{de}(I^{LR}_{real})) + \log(1 - D_{de}(I^{LR}_{de})) + a \cdot \left\| I^{LR}_{real} - I^{LR}_{de} \right\|_1, \tag{B5}$$

where D_{de} (utilizing the same structure as the PathGAN [14]) is a discriminator to distinguish I^{LR}_{de} from I^{LR}_{real}, and a is a weight hyperparameter.

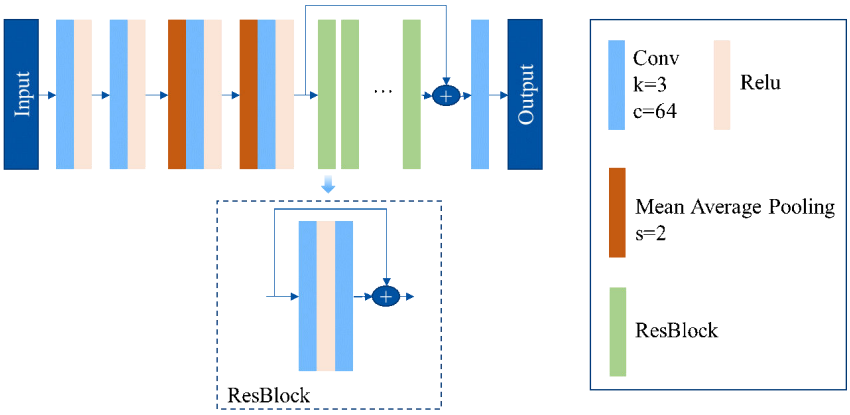


Figure B1 The network structure of G_{de} network.

Appendix C Experimental results

Appendix C.1 Evaluation dataset

Our method can be trained on unpaired data sets with different spatial resolutions captured by different satellites. To verify the effectiveness of the proposed method, we construct two datasets: a synthetic unpaired SR dataset and a real unpaired SR dataset. More details of these datasets are described as follows.

Synthetic unpaired SR dataset: The experiments of synthetic unpaired SR are conducted based on the RRSSRD [16] dataset. The HR images of the RRSSRD dataset are acquired from WorldView-2 and GaoFen-2, with the spatial resolution of 0.5m/pixel and 0.8m/pixel. In the training phase, to acquire the synthetic LR dataset, we degrade the real HR images to LR images by adding noise and downsampling. The noise degradation is the additive white gaussian noise (AWGN) with the noise level $\sigma = 7.65$ and the downsampling is the nearest interpolation with scale factor 4. To train our method in an unpaired way, we randomly select about half of the HR images (2021 images) from the RRSSRD training data set as the real HR images. The remaining half of the HR images (2022 images) are degraded to noisy LR images that are used as real LR images. In the testing phase, we conduct experiments on the RRSSRD test set. The test LR images are degraded from test HR images using the same degradation used for the training phase. We denote the test sets as synthetic test sets.

Real unpaired SR dataset: In the training phase, we train our model on the unpaired real HR data and real LR data. For real HR data, we select images with a spatial resolution of 0.1−0.15m/pixel from the DOTA dataset [17] collected from the Google Earth, GF-2, and JL-1 satellites. For real LR data, we choose the Google Ref data with the spatial resolution of 0.6m/pixel from the RRSSRD dataset [16] captured from Google Earth. The real HR data has 544 images of various sizes, while the real LR data has 4045 images with the size of 480×480. To balance of the amount of training data, we randomly crop the HR image into 4500 patches with the size of 960×960. Examples of real HR and LR data sets can be seen in Figure C1. In the testing phase, we adopt the Google Ref data from the RRSSRD test set as our real LR test set. In the RRSSRD test set, the Google Ref images in #1 test set and #2 test set are the same, and the Google Ref images in #3 test set and #4 test set are the same as well. Therefore, we choose the #1 test set and #3 test set as our test data sets and denote them as Real #1 test set and Real #2 test set, respectively.

(a) Real HR images

(b) Real LR images

Figure C1 Examples of real HR and LR images.

Appendix C.2 Evaluation metrics

To evaluate the proposed method, a blind image quality analyzer called as the natural image quality evaluator (NIQE) [18] is first used as the evaluation metric on both synthetic and real unpaired SR dataset. The NIQE does not require the ground truth and only calculates the measurable deviations from statistical regularities, without being trained on human-rated distorted images. The lower score of the reconstructed image represents better visual result. In the synthetic SR dataset, we adopt the peak signal-to-noise ratio (PSNR) [19], structural similarity (SSIM) [19], learned perceptual image patch similarity (LPIPS) [20], and NIQE for evaluation. The PSNR and SSIM are the most commonly used evaluation metrics in image SR. Given the ground truth image I with N pixels and reconstructed image $\hat{I}$, the PSNR is expressed as follows,

$$\mathrm{PSNR} = 10 \cdot \log_{10} \frac{L^2}{\frac{1}{N}\sum_{i=0}^{N-1} \left\| I(i) - \hat{I}(i) \right\|^2}, \tag{C1}$$

where L represents the maximum pixel value of the image. The SSIM is used to measure the structural similarity between images:

$$\mathrm{SSIM} = \frac{(2\mu_I \mu_{\hat{I}} + C_1)(\sigma_{I\hat{I}} + C_2)}{({\mu_I}^2 + {\mu_{\hat{I}}}^2 + C_1)({\sigma_I}^2 + {\sigma_{\hat{I}}}^2 + C_2)}, \tag{C2}$$

where μ and σ represent the mean and standard deviation of the image intensity, $C_1 = (k_1 L)^2$ and $C_2 = (k_2 L)^2$ are constants to avoid the denominator being 0 and maintain stability ($k_1 << 1$, $k_2 << 1$). The higher the scores of PSNR and SSIM, the better the reconstruction performance can be achieved. The LPIPS is a learned metric to measure the perceptual similarity between recovered and ground truth images using a pretrained deep network:

$$\mathrm{LPIPS} = \sum_l \frac{1}{H_l W_l} \sum_{h,w} \left\| w_l \odot (f^l_{h,w} - \hat{f}^l_{h,w}) \right\|_2^2, \tag{C3}$$

where H_l and W_l are the height and width of the features of l-th layer, $f^l_{h,w}$ and $\hat{f}^l_{h,w}$ represent the features of the corresponding I and $\hat{I}$ of the l-th layer at location (h, w). The reconstruction images with lower scores of LPIPS mean better image quality.

Appendix C.3 Implementation details

In our experiments, we adopt data augmentation on training samples, including the rotation of 90°, 180°, 270°, and horizontally flipping. In the training process of the domain translation, both the real LR images and bicubic downsampled images are cropped to the size of 64×64. In Eq. (1), the parameters m and n are set to 5 and 4, respectively. The parameters of Eq. (B1) are set to $\lambda_1 = \lambda_2 = 10$, $\lambda_3 = 1$. For the training of SR process, the patch sizes of pseudo-LR images and real LR images are set as 48×48 and the patch size of real HR images is 192×192. We set the parameters of Eq. (3) as $a = 1$ and the parameters of Eq. (B3) are set as $\omega_1 = 10, \omega_2 = 0.5, \omega_3 = 1, \omega_4 = 0.1$. All experiments are conducted based on Pytorch with two GTX-1080 Ti graphics processing

Figure C2 Experimental results on synthetic test sets. The results of the PSNR/SSIM/LPIPS of each SR image are presented. The best result is in bold. The second-best result is underlined.

units. For both domain translation and SR process, the training stops when the epoch is 500. The learning rate is chosen to be 1×10e-4 and the batch size is set as 16. The Adam optimization [21] is used to optimize the model parameters, where β_1 is 0.9 and β_2 is 0.999. Since the spatial resolution of real HR images is about four times higher than that of real LR images, the SR scale factor is selected to be 4 in our experiments.

Appendix C.4 Experiments on synthetic unpaired remote sensing images

In the synthetic unpaired remote sensing images, we adopt the PSNR, SSIM, NIQE, and LPIPS metrics on different SR methods. The compared CNN-based methods include the SRCNN [1], EDSR [22], our baseline model RRDBNet [10] and RCAN [23]. In

Table C1 Quantitative comparison with different methods on the synthetic test sets, where the degradation is AWGN with the noise level $\sigma = 7.65$. For PSNR and SSIM, a higher score indicates a better result. Whereas for NIQE and LPIPS, a lower score indicates a better result. In each row, the best result is labeled in bold and the second-best result is underlined.

Dataset	Metric	SRCNN	EDSR	RRDBNet	ESRGAN	ZSSR	RCAN	CinCGAN	BSRGAN	DASR	CPWSNN
Synthetic #1 Test	PSNR	24.69	26.81	26.83	24.65	26.45	25.88	26.74	26.78	27.22	**28.09**
	SSIM	0.628	0.696	0.682	0.556	0.657	0.629	0.735	0.710	0.641	**0.739**
	NIQE	9.770	8.487	8.939	5.528	9.007	8.671	6.744	4.879	3.761	**3.720**
	LPIPS	0.522	0.546	0.542	0.503	0.674	0.583	0.433	0.321	0.337	**0.254**
Synthetic #2 Test	PSNR	25.56	26.19	25.70	22.88	26.05	24.94	26.62	26.48	26.34	**27.83**
	SSIM	0.624	0.661	0.636	0.504	0.629	0.573	**0.715**	0.661	0.584	0.700
	NIQE	11.85	8.671	8.728	6.364	9.354	9.204	6.801	4.830	4.179	**3.647**
	LPIPS	0.754	0.625	0.619	0.584	0.746	0.666	0.456	0.407	0.374	**0.295**
Synthetic #3 Test	PSNR	23.48	25.10	24.71	23.34	25.04	24.36	25.39	25.54	25.52	**26.12**
	SSIM	0.568	0.633	0.609	0.499	0.607	0.575	0.658	0.649	0.581	**0.664**
	NIQE	9.798	9.082	9.148	5.778	8.886	9.135	6.791	4.238	3.587	**3.610**
	LPIPS	0.637	0.654	0.639	0.497	0.766	0.661	0.494	0.340	0.335	**0.274**
Synthetic #4 Test	PSNR	24.96	26.62	26.18	22.75	26.36	25.26	26.15	26.64	26.41	**28.11**
	SSIM	0.593	0.652	0.631	0.487	0.623	0.562	0.689	0.649	0.564	**0.697**
	NIQE	9.820	9.492	9.470	7.129	9.720	10.218	7.055	4.820	4.530	**3.983**
	LPIPS	0.605	0.635	0.627	0.596	0.747	0.664	0.459	0.425	0.407	**0.311**

addition, we choose the ESRGAN [10] as the typical GAN-based method for the comparison. One blind SR method, named BSRGAN [24], and an unsupervised SR method, named as ZSSR [25] are also included in the comparison. Furthermore, we select two unpaired natural image SR methods for the comparison, named as the CinCGAN [26] and DASR [27]. The HR images of training data used by the compared methods are the same. The degradation of CNN-based SR method and GAN-based method is bicubic. The blind SR method BSRGAN is trained with multiple degradation in the same experimental setting of BSRGAN. The unpaired SR methods are optimized on the synthetic unpaired dataset as mentioned in subsection Appendix C.1. All methods are evaluated on the synthetic test sets.

As shown in Table C1, we quantitatively evaluated the SR results using four metrics, including PSNR, SSIM, NIQE, and LPIPS, where the best result is labeled in bold and the second best result is underlined. As can be seen, our CPWSNN achieves the highest scores on almost all metrics. Besides this, there are visual and quantitative comparisons in Figure C2. The CNN-based methods including the SRCNN, EDSR, RRDBNet, and RCAN cannot reconstruct the rich details in the images and caused many artifacts. The GAN-based method ESRGAN mistakenly treats noise as high-frequency detail information and amplifies t he noise and artifacts. The results of the ZSSR method suffer f rom s erious n oise. T he r esults o f t he B SRGAN a re b lurry a nd structural details are not well preserved. CinCGAN and DASR have relatively good visual effects, b ut c ause c olor d istortion. T he visual results of our CPWSNN are the closest to HR images and have the highest scores in the quantitative metrics.

To further validate the robustness of our CPWSNN, we conduct experiments under mixed degradations. Specifically, we degrade the synthetic test sets by Gaussian blur (kernel size is 3), AWGN (noise level $\sigma = 7.65$), and nearest downsampling with the scale of 4 [28, 29]. We directly test the synthetic mixed degradations test sets on the above-mentioned methods without finetuning. The quantitative results are shown in Table C2. Our CPWNN has comparable results on the synthetic test sets under mixed degradations, demonstrating that our proposed method has high robustness.

Table C2 Quantitative comparison with different methods on the synthetic test sets under mixed degradations, where the degradations are Gaussian blur with kernel size = 3, and AWGN with the noise level $\sigma = 7.65$, and nearest downsampling with the scale of 4.

Dataset	Metric	SRCNN	EDSR	RRDBNet	ESRGAN	ZSSR	RCAN	CinCGAN	BSRGAN	DASR	CPWSNN
Synthetic #2 Test	PSNR	**27.38**	27.20	27.10	24.86	26.55	27.07	26.11	26.29	27.21	27.12
	SSIM	0.682	0.636	0.626	0.523	0.612	0.625	0.666	0.680	0.654	**0.690**
Synthetic #3 Test	PSNR	27.46	27.18	26.95	23.74	26.56	27.08	26.47	26.37	27.26	**27.50**
	SSIM	0.657	0.613	0.602	0.496	0.591	0.605	0.651	0.642	0.629	**0.668**
Synthetic #4 Test	PSNR	**26.51**	26.33	26.24	24.14	25.83	26.23	25.28	25.36	26.37	26.19
	SSIM	0.637	0.593	0.584	0.488	0.579	0.584	0.627	0.624	0.611	**0.639**
Synthetic #1 Test	PSNR	28.07	27.62	27.35	23.69	27.02	27.53	26.84	26.79	27.87	**28.10**
	SSIM	0.659	0.607	0.597	0.485	0.590	0.601	0.653	0.639	0.631	**0.667**

Appendix C.5 Experiments on real unpaired remote sensing images

In these experiments, all the compared methods are trained on the same dataset. To test the above CNN-based methods (SRCNN [1], EDSR [22], RRDBNet [10], RCAN [23]) and GAN-based method (ESRGAN [10]), we utilize the bicubic method to degrade the real HR data to the LR data. The blind SR method, named as BSRGAN [24], is trained with multiple degradation in the same experimental setting of BSRGAN. The unpaired SR methods (CinCGAN [26] and DASR [27]) are all trained on the real unpaired dataset mentioned in Appendix C.1. Since the real LR images have no corresponding ground truth images, we only utilize the NIQE metric to evaluate the SR performance, where the results are shown in Table C3. As can be observed, our CPWSNN is

Table C3 NIQE metric comparison of different methods on real test sets. A lower score indicates a better result. The best results are in bold and the sencond-best results are underlined.

DATASET	SRCNN	EDSR	RRDBNet	ESRGAN	ZSSR	RCAN	CinCGAN	BSRGAN	DASR	CPWSNN
Real #1 Test	8.047	4.792	7.075	5.278	7.396	5.290	6.034	<u>4.290</u>	6.447	**3.895**
Real #2 Test	9.206	5.691	7.775	<u>5.128</u>	7.434	6.161	8.075	5.483	6.943	**4.114**

superior to all the other methods on the NIQE metric. Visual comparisons of different methods are illustrated in Figure C3. As can be seen, the SR results of the CNN-based methods, which are SRCNN, EDSR, RRDBNet, and RCAN, cannot produce much detail and suffer from blur. The GAN-based method ESRGAN has a better visual result but amplifies blocky artifacts. The results of the unsupervised method ZSSR has certain suppression effects on blocky artifacts, while the textures and details of the SR results are not obvious. CinCGAN has similar results with the CNN-based methods both in visual effects and the NIQE metric. The results of the BSRGAN are overly smooth. The artifacts in DASR results is more serious. Compared with other methods, the proposed CPWSNN has the best visual effects. For example, the edges and textures information are richer and the objects can be seen more clearly in CPWSNN SR results.

Figure C3 Experimental results on Real test sets consist of real LR remote sensing images. The results of the NIQE of each SR image are presented. A lower score indicates better results.

Appendix C.6 Ablation studies

To further verify the effects of each loss to our method, we conducted ablation studies, including the perceptual loss in domain translation, the degradation consistency loss, and edge retention loss in SR training process.

Appendix C.6.1 *Studies of perceptual loss in domain translation*

We conduct experiments with and without perceptual loss in the domain translation training process to validate the content-preserving effectiveness of the perceptual loss. Visual results are shown in Figure C4. It is obvious that the objects and color information of pseudo-LR images generated by domain translation without perceptual loss is heavily changed. We can figure out that the perceptual loss has the strong ability to prevent objects deformation and color variation. Table C4 and Table C5 show the quantitative comparison on SR results with respect to the use of the perceptual loss in domain translation based on synthetic LR images and real LR images, respectively. As can be seen in Table C4 and Table C5, the performance is improved by adding the perceptual loss.

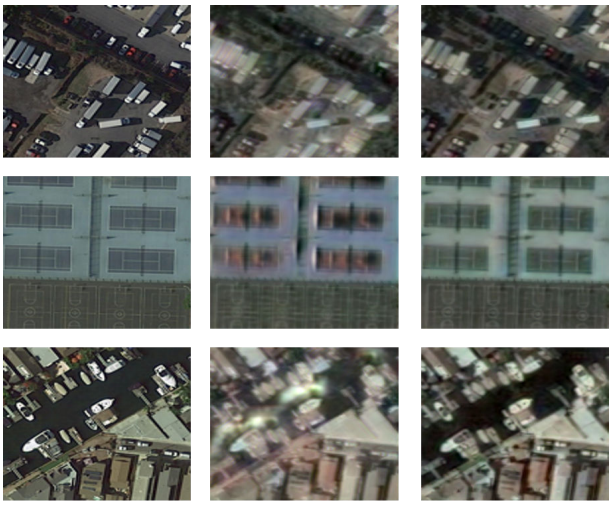


Figure C4 Visual effect of the perceptual loss in domain translation. (a) Real HR images (zoom in for best view). (b) Pseudo-LR images generated by domain translation without perceptual loss. (c) Pseudo-LR images generated by domain translation with the perceptual loss.

Table C4 Quantitative comparison on SR results with respect to the use of the perceptual loss in domain translation based on synthetic test sets. The best results are in bold.

Method	Synthetic #1 Test PSNR/SSIM/LPIPS	Synthetic #2 Test PSNR/SSIM/LPIPS	Synthetic #3 Test PSNR/SSIM/LPIPS	Synthetic #4 Test PSNR/SSIM/LPIPS
Without perceptual loss	26.50/0.694/0.385	26.54/0.661/0.432	24.84/0.614/0.331	26.26/0.651/0.439
With perceptual loss	**28.09/0.739/0.247**	**27.83/0.701/0.295**	**26.21/0.664/0.266**	**28.11/0.696/0.313**

Table C5 NIQE metric comparison on SR results about the perceptual loss in the domain translation based on real test sets. The best results are in bold.

Method/Test set	Real #1 Test	Real #2 Test
Without perceptual loss	4.659	4.939
With perceptual loss	**3.895**	**4.114**

Appendix C.6.2 *Studies on edge retention loss and degradation consistency loss in SR training process*

We designed several experiments to study the effects of the edge retention loss and degradation consistency loss in the SR training process (without the edge retention loss and degradation consistency loss, without edge retention loss, and the full loss). Table C6 shows the quantitative comparisons of the SR network trained with different losses based on the synthetic #1 test set. The SR network trained with the full loss function has the best quantitative results in terms of the metrics of PSNR and LPIPS. Figure C5 shows the visual results on different loss functions on real test sets. The SR network trained without edge retention loss and degradation consistency loss generate artificial results with color variation and unreal objects. By adding the degradation consistency loss, the color variation issue can be suppressed but the SR results still have artificial edges. When the edge retention loss is further adopted, the generation of fake edges and textures can be prevented. Figure C5(d) indicates that the degradation consistency loss and edge retention loss can provide constraints for the SR network to avoid generating unreal objects, color variation, unreal edges, and other content distortion issues.

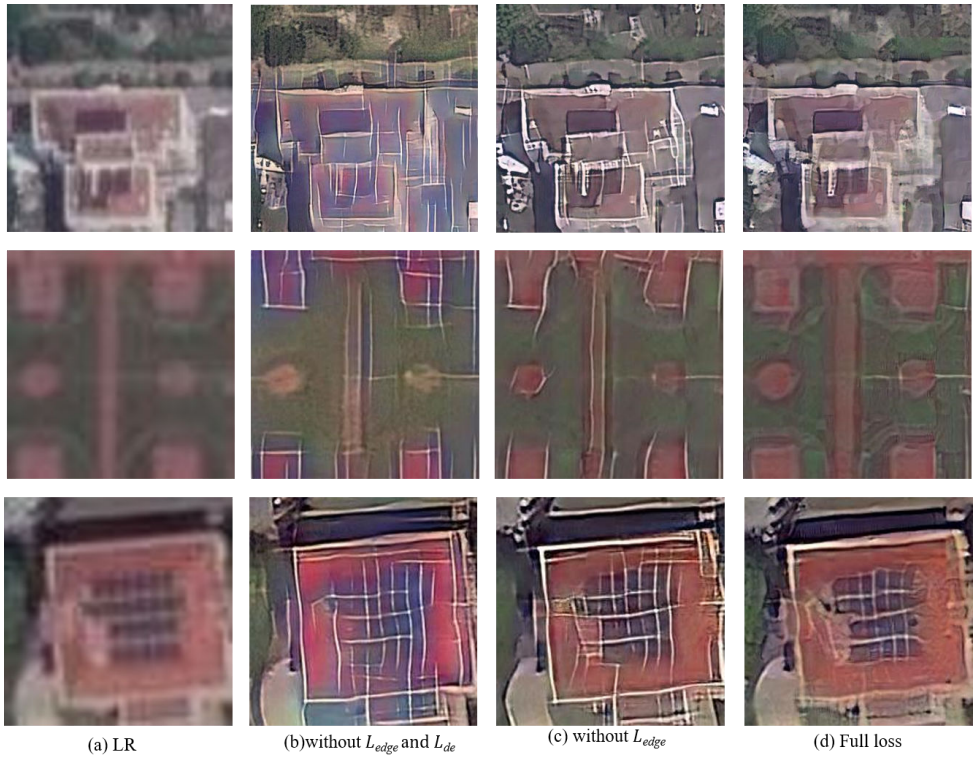


Figure C5 Visual comparison on SR network trained with different loss based on Real test sets.

Table C6 Quantitative comparison of SR network trained with different loss based on the real #1 test set.

MAE loss and adversarial loss	Edge retention loss	Degradation consistency loss	PSNR/SSIM/LPIPS
✓	×	×	21.28/0.601/0.432
✓	×	✓	23.58/0.593/0.385
✓	✓	✓	**28.09/0.739/0.254**

Appendix D Conclusion

We presented a content-preserving weak supervision neural network to reconstruct SR based on unpaired remote sensing data sets. First, we adopted the domain translation to synthesize the pseudo-LR images from real HR images. Then, the pseudo-LR images and real HR images provided pixel wise supervision for the SR network. To preserve the contents of the synthesized the pseudo-LR images, we used the perceptual loss to constrain the domain translation process. Furthermore, to avoid fake edges and unreal objects in the SR results, we proposed the edge retention loss and degradation consistency loss to constrain the SR network training. Experimental results on a synthetic test set and real test set demonstrated that the proposed unpaired remote sensing images SR method achieved competitive visual results compared with paired and unpaired SR methods.

References

1 Dong C, Loy C C, He K M, et al. Image super-resolution using deep convolutional networks. IEEE Trans Pattern Anal Mach Intell, 2015, 38: 295-307
2 Lei S, Shi Z W, Zou Z X. Super-resolution for remote sensing images via local–global combined network. IEEE Geosci Remote Sens Lett, 2017, 14: 1243-1247
3 Jiang K, Wang Z Y, Yi P, et al. Deep distillation recursive network for remote sensing imagery super-resolution. Remote Sens, 2018, 10: 1700
4 Lu T, Wang J M, Zhang Y D, et al.Satellite image super-resolution via multi-scale residual deep neural network. Remote Sens, 2019, 11: 1588
5 Pan Z X, Ma W, Guo J Y, et al. Super-resolution of single remote sensing image based on residual dense backprojection networks. IEEE Trans on Geosci Remote Sens, 2019, 57: 7918-7933
6 Zhang D Y, Shao J, Li X Y, et al. Remote sensing image super-resolution via mixed high-order attention network. IEEE Trans on Geosci Remote Sens, 2020, 59: 5183-5196
7 Huang G, Liu Z, Van D M L, et al. Densely connected convolutional networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017. 4700-4708
8 Dong X Y, Sun X, Jia X P, et al. Remote sensing image super-resolution using novel dense-sampling networks. IEEE Trans on Geosci Remote Sens, 2020, 59: 1618-1633
9 Ledig C, Theis L, Huszár F, 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, 2017. 4681-4690
10 Wang X T, Yu K, Wu S X, et al. Esrgan: Enhanced super-resolution generative adversarial networks. In: Proceedings of the European Conference on Computer Vision (ECCV) workshops, 2018. 0-0

11 Zhang W L, Liu Y H, Dong C, et al. Ranksrgan: Generative adversarial networks with ranker for image super-resolution. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019. 3096-3105
12 Jiang K, Wang Z Y, Yi P, et al. Edge-enhanced GAN for remote sensing image superresolution. IEEE Trans on Geosci Remote Sens, 2019, 57: 5799-5812
13 Lei S, Shi Z W, Zou Z X. Coupled adversarial training for remote sensing image super-resolution. IEEE Trans on Geosci Remote Sens, 2019, 58: 3633-3643
14 Isola P, Zhu J Y, Zhou T H, et al. Image-to-image translation with conditional adversarial networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017. 1125-1134
15 Zhu J Y, Park T, Isola P, et al. Unpaired image-to-image translation using cycle-consistent adversarial networks. In: Proceedings of the IEEE International Conference on Computer Vision, 2017. 2223-2232
16 Dong R M, Zhang L X , Fu H H. RRSGAN: Reference-Based Super-Resolution for Remote Sensing Image. IEEE Trans on Geosci Remote Sens, 2021: 1-17
17 Xia G S, Bai X, Ding J, et al. DOTA: A large-scale dataset for object detection in aerial images. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018. 3974-3983
18 Mittal A, Soundararajan R , Bovik A C. Making a "completely blind" image quality analyzer. IEEE Signal Process Lett, 2012, 20: 209-212
19 Wang Z, Bovik A C, Sheikh H R, et al. Image quality assessment: from error visibility to structural similarity. IEEE Trans on Image Process, 2004, 13: 600-612
20 Zhang R, Isola P, Efros A A, et al. The unreasonable effectiveness o f d eep f eatures a s a p erceptual m etric. I n: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018. 586-595
21 Kingma D P , Ba J. Adam: A method for stochastic optimization. 2014, arXiv: 1412.6980
22 Lim B, Son S, Kim H, et al. Enhanced deep residual networks for single image super-resolution. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2017. 136-144
23 Zhang Y L, Li K P, Li K, et al. Image super-resolution using very deep residual channel attention networks. In: Proceedings of the European Conference on Computer Vision, 2018. 286-301
24 Zhang K, Liang J, Van Gool L, et al. Designing a practical degradation model for deep blind image super-resolution. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021. 4791-4800
25 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, 2018. 3118-3126
26 Yuan Y, Liu S Y, Zhang J W, et al. Unsupervised image super-resolution using cycle-in-cycle generative adversarial networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2018. 701-710
27 Wei Y, Gu S, Li Y, et al. Unsupervised real-world image super resolution via domain-distance aware training. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021. 13385-13394
28 Zhang J, Lu S, Zhan F, et al. Blind image super-resolution via contrastive representation learning. 2021, arXiv:2107.00708
29 Zhang N, Wang Y, Zhang X, et al. A multi-degradation aided method for unsupervised remote sensing image super resolution with convolution neural networks. IEEE Trans on Geosci and Remote Sens, 2020, 60: 1-14