IEICE TRANSACTIONS

on Fundamentals of Electronics, Communications and Computer Sciences

DOI:10.1587/transfun.2024EAL2050

Publicized:2024/10/18

A PUBLICATION OF THE ENGINEERING SCIENCES SOCIETY

The Institute of Electronics, Information and Communication Engineers

Kikai-Shinko-Kaikan Bldg., 5-8, Shibakoen 3 chome, Minato-ku, TOKYO, 105-0011 JAPAN

LETTER

Progressive Multi-Scale Learning for Remote Sensing Image Super-Resolution with Residual Prior

Qiuyu XU†, Kanghui ZHAO†, Tao LU*†, Zhongyuan WANG††, *and* Ruimin HU††,

SUMMARY Global contextual information and spatial structural details are pivotal elements in the context of super-resolution (SR) reconstruction for remote sensing images. Therefore how to generate rich contextual semantic information and accurate spatial structure information simultaneously is a key challenge for remote sensing image SR. In this paper, we propose a novel progressive multi-scale learning strategy based on residual prior to solve the remote sensing image SR problem. In particular, we propose a novel progressive up-down mapping unit (PUMU) that asymptotically maps the input low-dimensional vectors into a high-dimensional space to learn global context information, which avoids loss of global information. Subsequently, we suggest introducing a novel method of explicitly mining spatial structure information, called residual prior (RP), which can help the proposed model to achieve spatial-structure-preserving SR. We have conducted extensive experiments on two public datasets including UCMerced and PatternNet, and the experimental results demonstrate the effectiveness of the proposed method.

key words: *remote sensing image super-resolution, progressive multi-scale learning, residual prior*

1. Introduction

Remote sensing high-resolution (HR) image is an important prerequisite for earth observation techniques and is widely used for high level vision tasks such as remote sensing segmentation and classification. However, due to the long imaging distance and the influence and limitation of the equipment, it results in a low resolution (LR) of the images captured, which cannot meet the application of the actual scene. Single image super-resolution (SISR) [1] technique can infer the corresponding HR image from the observed LR image by relying on the a prior knowledge, which can overcome the limit of the optical imaging system and provide better input images for the subsequent detection and recognition, etc. Therefore, the study of reconstructing high-quality satellite images by SR techniques is of great importance.

In recent years, deep learning has been widely introduced into the field of remote sensing image super-resolution (SR), such as RCAN [2], SAN [3], HAN [4], MHAN [5], CTNet [6], HSENet [7], MEN [8], and SPE [9]. Compared with SR methods based on interpolation and reconstruction, deep learning relies on powerful implicit feature representations and automatically constructs high-level representations of the original inputs to significantly improve the performance of SR reconstruction. However most methods tend to design complex network structures to learn LR to HR mappings directly, ignoring global contextual and spatial structure information.

Contextual information is a critical factor in SR of remote sensing image. Considering that multi-scale networks can expand the receptive field and thus capture finer content, i.e., contextual information, a series of multi-scale strategies have been applied to SR.Expanding the receptive field can be achieved through a straightforward and effective approach, namely, the utilization of multi-scale convolution. In a study by Wang *et al.* [8], they introduced a Multi-scale Enhancement Network (MEN) designed to harness the multi-scale features inherent in remote sensing images, thereby improving the network's reconstruction capability. Specifically, a fusion of convolutional layers employing multiple kernel sizes is employed to enhance the extraction of multi-scale features.However, when the input image size is fixed, its multiscale feature learning is limited. Another proven way is to use a U-Net-like encoding-decoding structure for image reconstruction. Gao *et al.* [10] processed the input images in two different downscaled spaces and designed a multi-scale residual network based on residual blocks in the downscaled space. However, it is well known that more spatial information is lost as the information flow goes deeper in the image coding process, which is unacceptable for SR tasks. In view of this, it is necessary to develop a novel multi-scale learning strategy that can simultaneously learn global contextual information without losing spatial information.

The recovery of spatial structure information is another key factor in the SR task for remotely sensed images. The spatial structure cares more about the overall image contour or edge details. Some previous researches such as RCAN [2], SAN [3], HAN [4], and MHAN [5] have achieved some success with SR methods based on attention mechanism. However, the images reconstructed by these methods based on prior representations of implicit spatial structures tend to be too smooth, and as the network deepens, there is a substantial loss of information and the training error accumulates. Thus explicitly mining the a prior information about the spatial structure is another challenge for SR.

To address the above problems, we propose a novel multi-scale learning strategy to learn image context global features without losing spatial structure information. Furthermore, we incorporate an image prior that effectively preserves the inherent structure of the image. This image prior is capable of explicitly learning a representation that

†The authors are with the Hubei Key Laboratory of Intelligent Robot, Wuhan Institute of Technology, Wuhan 430073, China.

††The authors are with NERCMS, School of Computer Science, Wuhan University, Wuhan 430079, China

†(Kanghui Zhao is co-first author, Corresponding author: Tao Lu(lutxyl@gmail.com)).

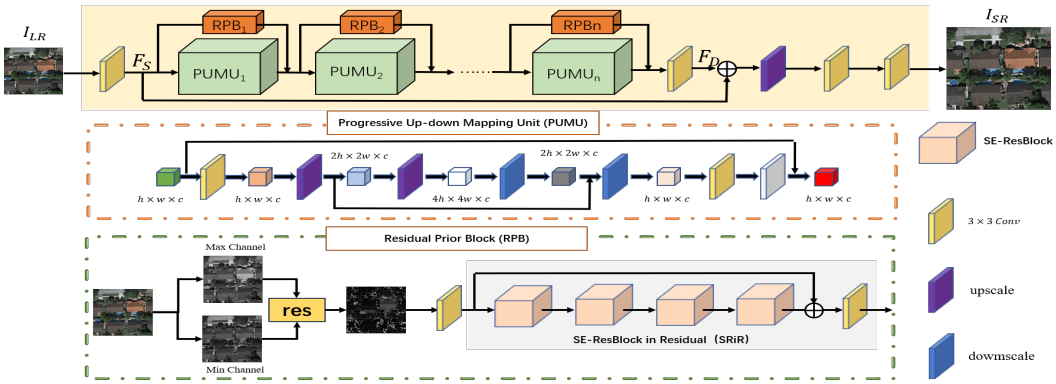

Fig. 1 The outlined framework: I_{LR} denotes the input, and I_{SR} signifies the ultimate super-resolved output in our proposed methodology.

guides the model towards high-quality image reconstruction. Specifically, we propose a progressive up-down mapping unit (PUMU) that progressively maps the input low-dimensional vectors from a low-dimensional space to a high-dimensional space. We attain the goal of broadening the receptive field without modifying the convolutional kernel size, and this remains unaffected by the dimensions of the input image. The designed PUMU avoids the information loss caused by the encoder, and at the same time can learn multi-scale global context information. Then, we explore another novel new paradigm for explicit expression of image prior, which we call residual prior (RP). RP does not rely on existing detection operators such as edges and gradients, and it is the residual result of the maximum and minimum channel values of an image, computed without any additional parameters. Therefore, we apply RP to the remote sensing image SR task and propose a residual prior block (RPB), which focuses more on learning spatial structural information and assists in generating high-quality HR images with clear and accurate structure. In summary, our contributions can be summarised as follows:

(1)we propose a novel multi-scale progressive learning strategy to learn fine-grained global contextual information by the designed PUMU that progressively maps images from a low-dimensional space to a multi-scale high-dimensional space. Subsequently, we introduce an explicit new paradigm of image prior representation, RP, and design RPB. We can ensure that our model mines the image spatial structure information without introducing additional parameter counts and computation.

(2) we construct a progressive multi-scale remote sensing image SR model based on residual prior bootstrapping. We perform a large number of experiments in UCMerced and PatternNet, and the experimental results prove that our method due to the existing state-of-the-art SR methods for remote sensing images.

2. PROPOSED METHOD

2.1 The whole Architecture of Proposed Network

We show the overall framework diagram of the proposed method in detail in Fig.1. We denote the input LR image as I_{LR} and the output image as I_{SR}. The aim of our model is to reconstruct the image I_{SR} from the input I_{LR} in an end-to-end manner. More specifically, the model can be divided into three components: a 3×3 convolutional layer for extracting shallow features from the input I_{LR}, followed by n PUMUs and their embedded RPBs for deep image feature representation, and finally an image reconstruction module consisting of an upsampling module and two 3×3 convolutional layers for reconstructing the output I_{SR}. Given a low-resolution input I_{LR}, we obtain initial shallow features F_S by applying a 3×3 convolutional layer

$$F_S = Conv_{3\times3}(I_{LR}), \tag{1}$$

where $Conv_{3\times3}$ denotes the 3×3 convolutional layer used for shallow feature extraction, F_S denotes shallow feature. Subsequently shallow feature F_S are fed into m PUMUs and RPBs for deep feature representation. The second part of the model network is multiple PUMUs and RPBs, which are stacked together by serial connections and embedding to capture accurate images feature step by step. This process can be represented as

$$F_D = Conv_{3\times3}(D_n(F_s)), \tag{2}$$

where D_n represents n PUMUs and embedded RPBs, F_D represents deep features. Finally, the SR image I_{SR} is generated by reconstruction as follows

$$I_{SR} = Conv_{3\times3}(Conv_{3\times3}(UP(F_S \oplus F_D))), \tag{3}$$

where UP denotes upsampling operation, and $\oplus$ represents element-level addition. $\{I^i_{LR}, I^i_{HR}\}_{i=1}^{N}$ is a given set of training sets,Incorporating N low-resolution images (I^i_{LR}) and their respective high-resolution counterparts (I^i_{HR}), the model training aims to minimize the L_1 loss function.

$$\hat{\theta} = argmin \frac{1}{N} \sum_{i=1}^{N} \left\| F\left(I^i_{SR}\right) - I^i_{HR} \right\|_1, \tag{4}$$

where θ represents the parameter set of the model, F represents the model function, and $\|.\|_1$ represents the L_1 loss

function.

2.2 Progressive Up-down Mapping Unit

The PUMU design is depicted in Fig.1, showcasing its simplicity and effectiveness. The input feature undergoes progressive upsampling, transitioning from low to high dimensions, facilitating the extraction of multi-scale features from the image.Then, we carry out progressive downsampling of these features and reduce them to the initial size as the output characteristics of PUMU. Different from the above methods.In our multi-scale strategy, LR images are mapped from low-dimensional to high-dimensional space, enabling multi-scale feature extraction within the expanded dimensional space of the image. This ensures the preservation of structure and texture details in the reduced-dimensional space while retaining the original multi-scale structure and texture features, contributing to the robustness of image SR reconstruction.

2.3 Residual Prior Block

Another important task of remote sensing image SR is to recover spatial structure texture information, which is more concerned with the overall image contour or edge details. The single scale network shows a great advantage in spatial information generation. However this implicit a prior mining is limited for spatial structure information recovery. To solve this problem, we suggest to introduce explicit a prior expressions. We introduce a residual prior (RP). The RP (Residual Prior) is independent of existing detection modalities like edges and gradients; it is computed based on the variability between the highest and lowest values within the image channels.As a result, it avoids introducing additional parameters and computational complexity.In addition, we design a residual prior block (RPB) to help RP learn spatial structural information, which helps to generate high-quality HR images with clear and accurate structure.

In Figure 1, it is evident that the RP image exclusively retains background details, delineated by the distinction between the highest and lowest channel intensities.It filters out low-frequency information from the image, but retains edges and textures, such as the outline of the house in the figure. Thus, given an input LR image I_{LR}, the RP of I_{LR} can be defined as

$$RP\,(LR) = \max_{c\in r,g,b} I_{LR} - \min_{c\in r,g,b} I_{LR}. \tag{5}$$

RP can extract target structures and textures more comprehensively and accurately.Hence, we incorporate RP into the image super-resolution (SR) model. To capture high-dimensional features of RP, we introduce a Residual Prior Block (RPB), as depicted in Fig. 1.The initial feature map obtained by convolutional layer first. In addition, to diminish noise within the initial features and enhance the semantic richness of these features, we directly use SRiR [11] to extract the deep features of the RP.

3. Experimental results and analysis

3.1 Dataset and Implementation Details

UCMerced [12] is a public dataset released by the UC Merced Computer Vision Laboratory for remote sensing image scene classification. The images come from manually extracted large images of urban areas across the country from the National Map Urban Area Imagery of the United States Geological Survey. The dataset includes 21 categories, and each category has 100 images of size 256×256 pixels. We select 90 images from each category for training and validation, and 10 images for testing. Therefore, the training and validation images amount to 1890, while the test images amount to 210.

PatternNet [13] is a large-scale, high-resolution remote sensing dataset collected for remote sensing image retrieval. There are 38 classes, with each class containing 800 images of size 256×256 pixels. The images in PatternNet are of various US cities, collected from Google Earth images or through the Google Map API. For training and validation, we select 62 images from each class, while 8 images are reserved for testing. As a result, there are a total of 2356 training and validation images and 304 test images.

During the model training stage, we set batchsize to 32 and patchsize is 64. We selected four commonly used objective evaluation metrics to evaluate the performance of reconstruction, including:PSNR, SSIM [14], and VIF [15]. In terms of details, the network was optimised for the training process using the ADAM [16] optimiser, where the learning rate was set to 0.0002, $\beta_1 = 0.9$, $\beta_2 = 0.999$. Our model consists of 10 PUMUs and corresponding RPBs, with the number of feature maps set to 64. The implementation of our model is based on the PyTorch framework and runs on a GPU with NVIDIA GTX 1080. For dataset processing, we use 4× bicubic interpolation to downsample HR images with resolution of 256×256 to LR images with resolution of 64×64.

3.2 Ablation Study

Verify the effect of RP and RPB. To substantiate the role of RP and RPB, a series of experiments were conducted for verification purposes. RPB was intentionally excluded from the proposed methodology, and the corresponding experimental outcomes are presented in Table 1. The deduction drawn from the results is that the removal of RPB leads to a degradation in performance, resulting in a decrease in PSNR from 30.79 dB to 30.40 dB. In addition, we provide another ablation study removing the RP and employing only the SRiR directly on the features that should be performed to validate our affirmations. The results of which are shown in Table 1. When we remove the RP, the performance of the model shows a significant decline in PSNR values by 0.10 dB, which demonstrates the role of RP in our model. To dynamically portray the demonstration of RP and RPB's significance, we visually present the subjective effects of retaining and removing RP and RPB in Fig.2. We chose two typical images for demonstration purposes. We can see that when removing the RP and RPB there is a wrong structure and texture. In summary, after removing the RP and RPB, the reconstruction lost more details and appeared blurred, which further confirms the importance of the RP and RPB in local detail texture recovery. We further show the MSE

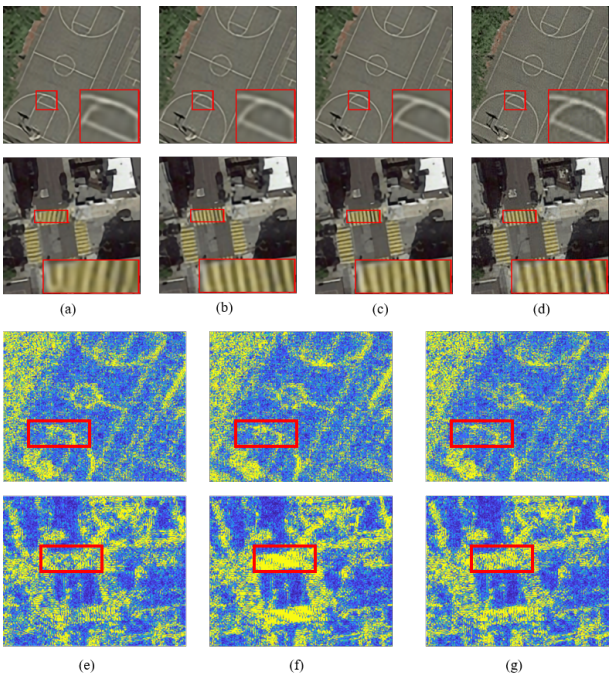

Fig. 2 Qualitative comparison of different settings of the proposed model. (a)Reconstruction result of removing RP. (b)Reconstruction result of removing RPB. (c) Reconstruction result of final model proposed. (d)HR. (e) MSE Map of removing RP. (f) MSE Map of removing RPB. (g) MSE Map of final model proposed. (Zoomed-in view to see more details.)

Table 1 VERIFY THE EFFECT OF RP and RPB. "OURS W/O RP" and "OURS W/O RPB" INDICATES THAT THE RP AND RPB IS REMOVED. The best results are **highlight**.

Methods	PSNR/dB	SSIM [14]	VIF [15]
Ours w/o RP	30.69	0.8211	0.4650
Ours w/o RPB	30.40	0.8157	0.4553
Ours	**30.79**	**0.8241**	**0.4698**

between SR and HR. We observed when we removed RP and RPB, satellite images with complex textures and dense objects complex textures and dense objects are easily contaminated by artifacts.

Verify the effect of batchsize and learning rate. To substantiate the role of batchsize and learning rate, a series of experiments were conducted for verification purposes, and the corresponding experimental outcomes are presented in Table 2 and Table 3. A smaller batchsize reduces memory footprint while meaning that gradient estimates are nosier with each update, which helps models jump out of local minimums and allows for training larger models or processing larger data sets with limited resources. However, too small a batchsize can lead to computational inefficiency and can also lead to unstable training processes. In contrast, a larger batchsize can improve computational efficiency, reduce noise from gradient estimation, and make the training process more stable, but may exceed memory limits and trap the model in local minima. Therefore, we set batchsize to 32 to achieve a compromise in terms of model training, computation, hardware requirements, and so on. Similarly, too

Table 2 Explore the impact of batchsize reconstruction results. The best results are **highlight**.

batchsize	PSNR/dB	SSIM [14]	VIF [15]
8	30.65	0.8210	0.4648
16	30.69	0.8212	0.4651
24	30.73	0.8226	0.4679
32	30.79	**0.8241**	0.4698
40	**30.80**	0.8240	**0.4701**

Table 3 Explore the impact of learning rate reconstruction results. The best results are **highlight**.

learning rate	PSNR/dB	SSIM [14]	VIF [15]
0.0001	30.70	0.8221	0.4668
0.0002	**30.79**	**0.8241**	**0.4698**
0.0003	30.75	0.8233	0.4689

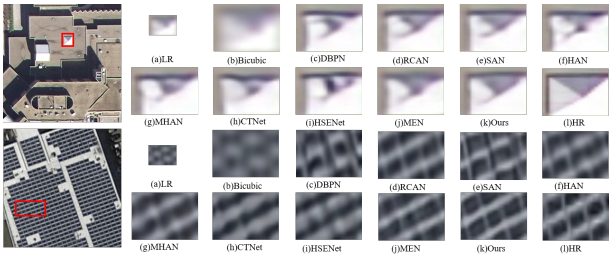

Fig. 3 In comparing our proposed method with eight other state-of-the-art approaches, we've emphasized critical areas in the image using red wireframes.

high learning rate may cause the model to skip the global minimum and fall into the local minimum during training. However, too low learning rate may cause the training process to be too slow or even stagnant. Proper learning rate is helpful for the model to find better solutions in the training process, thus improving the final performance of the model, so we choose 0.0002 as the final learning rate.

3.3 Compared with State-of-the-Arts

In this subsection, we have selected some state-of-the-art SR algorithms for comparison, including DBPN [17], RCAN [2], SAN [3], HAN [4], MHAN [5], CTNet [6], HSENet [7], and MEN [8]. For all SOTA models, we retrained and tested them through the same dataset using the official code provided by the authors.The numeric results are displayed in Table 4. As can be seen from the quantitative results, our method is optimal on both datasets. Specifically, on the UCMerced dataset and PatternNet dataset, our approach demonstrates a performance advantage of at least 0.07 dB and 0.11 dB over other methods, providing strong evidence for the effectiveness of our model.

As can be seen from Fig.3, although some texture details can be inferred from the reconstruction results of other SR algorithms, the lower feature utilization produces excessively smooth results, with artifacts at the edges of the resulting image and very blurred texture details. In contrast, our method can reconstruct images with more realistic texture details and fewer artifacts.

Table 4 Qualitative comparison of reconstruction results comparing different algorithms under UCMerced and PatternNet datasets with scale of 4×. The best results are **highlight**.

Method	Dataset	PSNR/dB	SSIM [14]	VIF [15]
Bicubic	UCMerced [12]	26.48	0.6866	0.3569
DBPN [17]		28.33	0.7762	0.4226
RCAN [2]		28.20	0.7719	0.4175
SAN [3]		28.26	0.7725	0.4200
HAN [4]		28.25	0.7737	0.4198
MHAN [5]		28.12	0.7673	0.4144
CTNet [6]		27.98	0.7601	0.4117
HSENet [7]		28.39	0.7759	0.4262
MEN [8]		28.16	0.7660	0.4163
Ours		**28.46**	**0.7807**	**0.4278**
Bicubic	PatternNet [13]	28.05	0.7362	0.3856
DBPN [17]		30.28	0.8175	0.4614
RCAN [2]		30.53	0.8187	0.4617
SAN [3]		30.53	0.8188	0.4617
HAN [4]		30.53	0.8187	0.4621
MHAN [5]		30.31	0.8153	0.4535
CTNet [6]		30.27	0.8120	0.4536
HSENet [7]		30.68	0.8233	0.4667
MEN [8]		30.45	0.8154	0.4588
Ours		**30.79**	**0.8241**	**0.4698**

4. Conclusion

In this letter, we propose a novel residual prior-guided progressive multi-scale feature learning method for remote sensing image SR. Firstly, we propose a novel progressive multiscale learning strategy that can help our model capture global multi-scale contextual information without losing spatial structure information. Subsequently, we suggest to introduce RP as a complement to the spatial structure information. RP is the difference between the maximum and minimum values of image channel and does not introduce additional parameters and computations. With the help of RP, our model can generate SR images with clearer structural texture. experimental and ablation studies on two publicly available datasets have shown that the proposed method exhibits state-of-the-art performance. However, our method can only perform high-quality reconstruction of single-frame images, but cannot reconstruct low-resolution satellite videos well. Therefore, in future research, we will focus on super-resolution reconstruction algorithms for video satellite application scenarios.

5. Acknowledgments

This work has been supported by the National Natural Science Foundation of China (62072350, 62171328, 62401410); the Central Government Guides Local Science and Technology Development Special Projects (ZYYD2022000021); the National Natural Science Foundation of Hubei (2023AFB158); the Enterprise Technology Innovation Project (2022012202015060); the Youths Science Foundation of Wuhan institute of technology (XJ2023000103).

References

[1] K. Zhao, T. Lu, J. Wang, Y. Zhang, J. Jiang, and Z. Xiong, "Hyper-laplacian prior for remote sensing image super-resolution," IEEE Transactions on Geoscience and Remote Sensing, vol.62, pp.1–14, 2024.

[2] Y. Zhang, K. Li, K. Li, L. Wang, B. Zhong, and Y. Fu, "Image super-resolution using very deep residual channel attention networks," Proceedings of the European conference on computer vision (ECCV), pp.286–301, 2018.

[3] T. Dai, J. Cai, Y. Zhang, S.T. Xia, and L. Zhang, "Second-order attention network for single image super-resolution," Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp.11065–11074, 2019.

[4] B. Niu, W. Wen, W. Ren, X. Zhang, L. Yang, S. Wang, K. Zhang, X. Cao, and H. Shen, "Single image super-resolution via a holistic attention network," Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XII 16, pp.191–207, Springer, 2020.

[5] D. Zhang, J. Shao, X. Li, and H.T. Shen, "Remote sensing image super-resolution via mixed high-order attention network," IEEE Transactions on Geoscience and Remote Sensing, vol.59, no.6, pp.5183–5196, 2020.

[6] S. Wang, T. Zhou, Y. Lu, and H. Di, "Contextual transformation network for lightweight remote-sensing image super-resolution," IEEE Transactions on Geoscience and Remote Sensing, vol.60, pp.1–13, 2022.

[7] S. Lei and Z. Shi, "Hybrid-scale self-similarity exploitation for remote sensing image super-resolution," IEEE Transactions on Geoscience and Remote Sensing, vol.60, pp.1–10, 2021.

[8] Y. Wang, Z. Shao, T. Lu, C. Wu, and J. Wang, "Remote sensing image super-resolution via multiscale enhancement network," IEEE Geoscience and Remote Sensing Letters, vol.20, pp.1–5, 2023.

[9] T. Lu, K. Zhao, Y. Wu, Z. Wang, and Y. Zhang, "Structure-texture parallel embedding for remote sensing image super-resolution," IEEE Geoscience and Remote Sensing Letters, vol.19, pp.1–5, 2022.

[10] S. Gao and X. Zhuang, "Multi-scale deep neural networks for real image super-resolution," Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops, pp.0–0, 2019.

[11] Q. Yi, J. Li, Q. Dai, F. Fang, G. Zhang, and T. Zeng, "Structure-preserving deraining with residue channel prior guidance," Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), pp.4238–4247, October 2021.

[12] Y. Yang and S. Newsam, "Bag-of-visual-words and spatial extensions for land-use classification," Proceedings of the 18th SIGSPATIAL international conference on advances in geographic information systems, pp.270–279, 2010.

[13] W. Zhou, S. Newsam, C. Li, and Z. Shao, "Patternnet: A benchmark dataset for performance evaluation of remote sensing image retrieval," ISPRS journal of photogrammetry and remote sensing, vol.145, pp.197–209, 2018.

[14] Z. Wang, A.C. Bovik, H.R. Sheikh, and E.P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," IEEE transactions on image processing, vol.13, no.4, pp.600–612, 2004.

[15] H.R. Sheikh and A.C. Bovik, "Image information and visual quality," IEEE Transactions on image processing, vol.15, no.2, pp.430–444, 2006.

[16] D.P. Kingma and J. Ba, "Adam: A method for stochastic optimization," arXiv preprint arXiv:1412.6980, 2014.

[17] M. Haris, G. Shakhnarovich, and N. Ukita, "Deep back-projection networks for super-resolution," Proceedings of the IEEE conference on computer vision and pattern recognition, pp.1664–1673, 2018.