IGARSS 2022 - 2022 IEEE International Geoscience and Remote Sensing Symposium | 978-1-6654-2792-0/22/$31.00 ©2022 IEEE | DOI: 10.1109/IGARSS46834.2022.9883316

LEARNING AN INTRINSIC GRAPH NEURAL NETWORK FOR SARTELLITE VIDEO SUPER-RESOLUTION

Yi Xiao[1], *Xin Su*[2], *Qiangqiang Yuan*[1], *Member IEEE*

[1]School of Geodesy and Geomatics, Wuhan University, Wuhan, China.
[2]School of Remote Sensing and Information Engineering, Wuhan University, Wuhan, China

ABSTRACT

Existing video super-resolution (VSR) methods usually merge the redundant temporal information along frames to achieve information enhancement, which naturally discards the spatial redundancy information. This paper proposes an intrinsic Graph Neural Network (GNN) framework for satellite VSR to fully explore the internal spatial prior while considering the temporal information in the video frame sequence. Firstly, a Multi-Scale Deformable convolution (MSD) is adopted to accurately model the spatial-temporal relationship between frames. Then, we search for *k*-nearest neighbors to construct the spatial graph and profoundly excavate the prior spatial information brought by patch recurrence. Finally, the spatial-temporal redundant information is integrated and complementary. Experiments on Jilin-1 satellite video demonstrate the effectiveness of our framework.

Index Terms— Satellite video, super-resolution, graph neural network, deep learning

1. INTRODUCTION

Super-resolution (SR) is a classical low-level vision task, which can be categorized into single-image super-resolution (SISR) and video super-resolution (VSR) [1]. SISR recovers a high-resolution (HR) image from its corresponding low-resolution (LR) image in the spatial domain, while VSR requires to exploit temporal information from multiple adjacent frames to reconstruct an HR target frame. Therefore, VSR is a more challenging ill-posed inverse problem than SISR since it is demanded to model the relationship of frames in the temporal dimension.

Nowadays, most of the deep learning-based VSR methods carry out VSR in the following three steps: firstly, the spatial-temporal relationship can be described by aligning frames to the target frame, then the temporal information is aggregated by fusing the aligned frame, and finally, the HR target frame is reconstructed through an up-sampling operation. The alignment methods are represented by optical flow and kernel-based methods (such as deformable convolution and non-local). The optical flow-based method achieves explicit alignment at the sub-pixel level by using the optical flows between the target and adjacent frames to warp their corresponding adjacent frames. The kernel-based method implicitly realizes alignment at the feature level. For example, the deformable convolution-based approach estimates a convolution kernel where the sampling grid is deformed for each position to capture the redundant information that is brought by pixel motion. Since the receptive field of convolution is limited, the non-local idea is proposed to search for the global similar relationship between pixels. Thanks to the elaborate modeling of temporal information, the current deep learning-based VSR method has progressed considerably against the traditional method. However, since more attention has been paid to the redundant information in the temporal domain, the intrinsic spatial redundant information of the target frame is naturally neglected. Learning the spatial redundancy can complement the temporal redundancy information and guide the network to converge to the optimal solution under the constraint of the spatial prior.

Many studies [2] indicate that there is internal patch-recurrence in a natural image, which means a patch can be discovered in another position in the image with multiple highly similar patches. This inspired us to utilize the redundant information in these similar patches to fuse a more informative patch in the spatial dimension. Similar patches may present at a position farther away from the target patch and are distributed discretely. Therefore, the GNN naturally fits the task of modeling patch-recurrence. Specifically, each patch can be abstracted as a node of the graph, and the edge describes the similarity between nodes. Finally, we aggregate these patches with the weight of the edge.

In satellite video, temporal information is difficult to be mined due to the scarcity of moving pixels in large and wide remote sensing images. In this case, the spatial prior of the target frame becomes particularly important. Besides, moving objects have various scales, which makes it more challenging to realize feature alignment. Therefore, we introduce a MSD alignment module to take into account multi-scale redundant information simultaneously. Finally, our network will deeply integrate temporal and spatial redundant information to achieve complementarity.

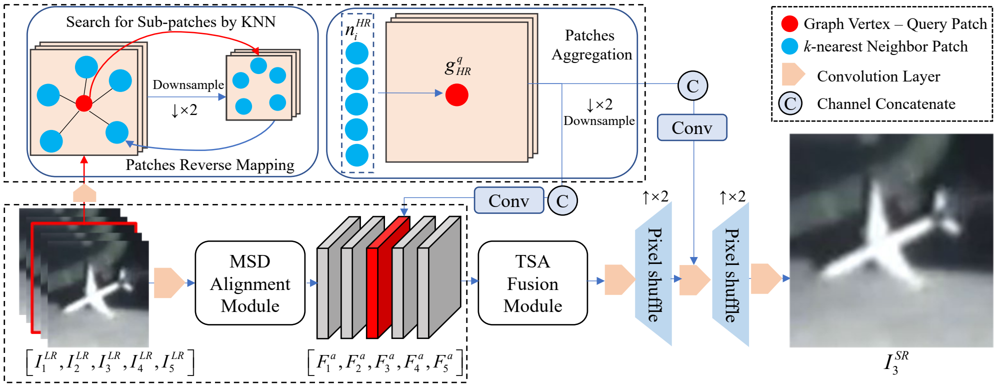

Fig. 1. The overall diagram of our intrinsic graph neural network.

2. METHODOLOGY

2.1. Overview

The overall diagram of our intrinsic graph neural network is shown in Fig. 1. Five LR consecutive frames $\left[I_1^{LR}, \cdots, I_5^{LR}\right]$ are the input. The intermediate frame I_3^{LR} is the target frame which need to be super-resolved (See red box), and the rest of the four frames are adjacent frames. Two branches are designed for spatial and temporal redundant information exploration.

In the spatial branch, a graph is firstly constructed on the feature maps of the target frame. We searched for the *k*-Nearest Neighbors (KNN) in the ×2 downsampling space $LR_{\downarrow 2}$ to form the nodes, and the similarity between nodes constitutes the edges. In another branch, we adopt the Multi-Scale Deformable convolution (MSD) alignment module proposed in our previous work [1] to realize feature alignment. As shown in Fig. 1., we conducted the spatial-temporal redundant information fusion in subsequent two × 2 up-sampling processes.

2.2. Spatial Branch

KNN Search. When searching *k*-nearest neighboring patches to query patch q^{LR}, Euclidean distance between the neighboring patch $n^{LR_{\downarrow 2}}$ and query patch was calculated to measure the similarity, that is:

$$dist(q^{LR}, n^{LR_{\downarrow 2}}) = \sqrt{\left\| q^{LR} - n^{LR_{\downarrow 2}} \right\|^2}, \tag{1}$$

and the top k patches in similarity are selected as the nodes. Here we search patches in × 2 bicubic downsampling features $LR_{\downarrow 2}$ to better learn the representation from LR patch to HR patch.

Nodes Reverse Mapping. After searching for k neighboring nodes with the same size as the query node in $LR_{\downarrow 2}$, we mapped the node positions back to the original LR space. The node size has also been increased by 2 times. These k HR patches $n_i^{HR}\left(i = 1:k\right)$ will be aggregated to obtain the HR patch corresponding to the query patch in latent HR feature space.

Patches Aggregation. The edges between HR nodes were used as the weight guide for aggregation. The aggregated patch is denoted as g_{HR}^q, here *HR* means that its size is twice of q^{LR}. A three-layer convolution $F(\cdot)$ was used to estimate the aggregation weight w_i, namely:

$$w_i = \exp\left(F\left(dist\left(q^{LR}, n_i^{LR_{\downarrow 2}}\right)\right)\right), \tag{2}$$

Finally, the aggregated patched are written as:

$$g_{HR}^q = \frac{1}{\delta}\sum_{i=1}^{k} w_i \cdot n_i^{HR} \tag{3}$$

2.3. Temporal Branch

MSD Alignment Module. We introduced the MSD Alignment Module to achieve implicit alignment at the feature level. All adjacent frames were aligned to the target frame. The details of the deformable convolution operation can be found in [1]. As shown in Fig. 1. The target frame feature (Marked red) and the downsampled HR features obtained in the spatial branch were merged to achieve the spatial-temporal redundant information fusion.

TSA Fusion Module. To ensure that the spatial information of the target frame is the dominant in the fusion process, we dynamically considered the contribution of different spatial-temporal information by introducing the TSA fusion module in EDVR. For details, please refer to [3].

3. EXPERIMENTAL RESULTS

Data Set. As in [1], 8 of the 10 original videos of the Jilin-1 satellite were cropped into 189 video clips with a size of 640×640, and each clip contains 100 consecutive frames to

Table. 1. The average PSNR/SSIM in six test video clips.

Method	Scene-1	Scene-2	Scene-3	Scene-4	Scene-5	Scene-6	Overall
Bicubic	29.71/0.8801	32.27/0.9252	28.59/0.8635	30.62/0.9009	32.47/0.9246	30.83/0.8989	30.75/0.8989
DUF	30.83/0.9310	32.69/0.9567	29.24/0.9155	31.60/0.9385	33.73/0.9644	31.48/0.9369	31.60/0.9405
TDAN	32.96/0.9392	36.11/0.9619	31.71/0.9254	34.07/0.9476	38.43/0.9726	34.17/0.9450	34.58/0.9485
Ours	**33.06/0.9393**	**36.27/0.9627**	**31.82/0.9262**	**34.42/0.9507**	**38.92/0.9749**	**34.44/0.9468**	**34.82/0.9501**

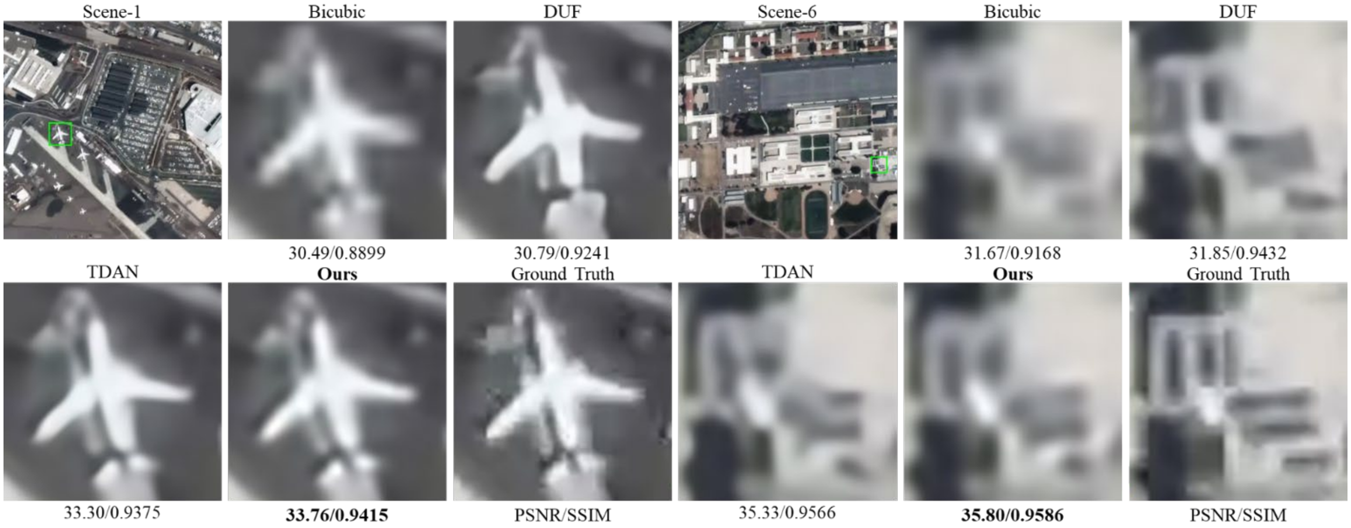

Fig. 2. Quantitative results in scene-1 and scene6.

form our training set. We used the *imresize* function in MATLAB to get LR video clips through the *Bicubic* kernel**.** Six scenes were randomly cropped to form our test set in the remaining two original videos.

Training details. This paper only conducts the ×4 SR. In the graph, *k* is set to 5, and the size of each query node is 3×3, the search window when searching for neighboring nodes is limited to 30 × 30. LR video frames as input were randomly cropped into patches with a size of 80×80. Image flipping and random rotation were also used for data augmentation. The min-batch is set to 1, and the learning rate is 1×10^{-5}. We adopted Adam as the optimizer and the $\mathcal{L}_1$ loss function to guide the optimization of our network.

Results. The quantitative results on the 6 test videos are shown in Table. 1. Our method is significantly ahead of the previous methods in PSNR/SSIM. The results illustrate the effectiveness of designing graph convolution branch to explore spatial redundancy information.

The qualitative results on scene-1 and scene-6 are shown in Fig. 2. In scene-1, our method gets the clearest airplane outline. DUF [4] has serious distortions, and TDAN [5] predicts sharp boundary. In scene-6, we pay attention to the boundary of the building on the ground, and we can see that only our method gets the closest result to the ground true. The rest of methods have artifacts and fuzzy. This demonstrates that spatial redundancy information provides valuable spatial priors for recovering realistic images.

4. CONCLUSION

In this paper, we propose a satellite VSR framework based on graph convolution. The network can fully excavate spatial-temporal redundant information and perform well on jilin-1 satellite.

5. REFERENCES

[1] Y. Xiao, X. Su, Q. Yuan, D. Liu, H. Shen, and L. Zhang, “Satellite video super-resolution via multiscale deformable convolution alignment and temporal grouping projection,” *IEEE Trans. on Geosci. Remote Sens*., early access, Sep. 6, 2021, doi: 10.1109/TGRS.2021.3107352.

[2] S. Zhou, J. Zhang, W. Zuo, and C. C. Loy, “Cross-scale internal graph neural network for image super-resolution,” 2020, *arXiv:2006.16673*.

[3] X. Wang, K. C. K. Chan, K. Yu, C. Dong, and C. C. Loy, “EDVR: Video restoration with enhanced deformable convolutional networks,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2019, pp. 1–10.

[4] Y. Jo, S. W. Oh, J. Kang, and S. J. Kim, “Deep video super-resolution network using dynamic upsampling filters without explicit motion compensation,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2018, pp. 3224–3232.Y. Tian,

[5] Y. Zhang, Y. Fu, and C. Xu, “TDAN: Temporally-deformable alignment network for video super-resolution,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 3360–3369.