IGARSS 2021 - 2021 IEEE International Geoscience and Remote Sensing Symposium | 978-1-6654-0369-6/21/$31.00 ©2021 IEEE | DOI: 10.1109/IGARSS47720.2021.9553281

A RECURRENT REFINEMENT NETWORK FOR SATELLITE 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

Deep learning-based methods have shown superior performance in VSR tasks. However, satellite video frames are characterized by large width, low resolution, and lack of features. Consequently, the conventional VSR method is not suitable for satellite video. In this paper, a recurrent refinement network is proposed. Considering that the vast majority of remote sensing images belong to the static background, a single-image SR (SISR) method is first used to obtain high-resolution features for a specific target frame. To further complement the missing details, the network learns the complementary information enhanced by an Encoder-Decoder structure from adjacent frames to refine the results of SISR. To measure the contribution of different adjacent frames to the recovery of the target frame, a temporal attention mechanism is introduced in the final fusion stage. The experiment on the video data of Jilin-1 demonstrates the effectiveness of our method.

Index Terms— Satellite video, super-resolution, deep learning

1. INTRODUCTION

Super-resolution (SR) is a low-level computer vision task, which aims to recover the corresponding high resolution (HR) image from the low resolution (LR) image [1]. SR can be divided into single-image SR (SISR), multi-image SR (MISR), and video SR (VSR). This paper focuses on the VSR method by fusing complementary information generated by subpixel motion between adjacent frames.

At present, most of the deep learning-based VSR method adopts the following procedure: sub-pixel alignment between frames, information fusion, and reconstruction. VESCPN [2] is the first end-to-end VSR method that simultaneously trains optical flow estimation and spatial-temporal information fusion. Xue et al. [3] proposed TOFlow, which used optical flow to estimate motion information between reference frames and target frame and then used the estimated optical flow to distort reference frames to achieve motion compensation. The optical flow-based method is called explicit motion compensation. However, the optical flow-based approach has a drawback: inaccurate optical flow estimation will lead to incorrect image distortion, and the frame misalignment gives the wrong information for subsequent fusions, resulting in poor reconstruction. Inspired by the deformable convolutional network, Wang et al. [4] proposed a VSR model with implicit motion compensation named EDVR. At first, the feature alignment is realized by introducing deformable convolution. Next, a spatial-temporal attention fusion module is used to fuse the aligned features. At last, 10 residual blocks and the subpixel convolution are used to get the final SR frame. Note that EDVR is a state-of-the-art method.

Compared with terrestrial video, satellite video can be more complicated, so the conventional VSR method cannot be applied directly. Firstly, satellite images have a vast width and contain more complex ground objects with multi scales. Secondly, the moving target is relatively small relative to the broad background and usually only occupies a small number of pixels in the image. Because of the low contrast and the pixel mixing at the edge of the object, it is difficult to distinguish the moving object from the background. Thirdly, due to the limitation of the resolution, remote sensing image lacks rich texture details, which makes it difficult to extract features. In a word, all the reasons mentioned above make satellite VSR a more challenging task.

Considering the characteristics of satellite video data and the drawbacks of the work mentioned above, we propose a recurrent refinement network. First of all, the vast majority of remote sensing images belong to the static background, which does not need complex MISR methods for reconstruction. Instead, we only need to learn the missing details from adjacent frames and add them to the results of SISR to refine the frame texture details. Secondly, in remote sensing images, the motion target is too small and the motion information is too little. The calculation cost of deformable convolution is too high, so it is not suitable. Thus, we use optical flow to capture the motion information between the target frame and reference frames. To prevent incorrect alignment, we do not distort the reference frame. Instead, the target frame, reference frame and their corresponding optical flow are all input into the network, so

Authorized licensed use limited to: Wuhan University. Downloaded on February 26,2022 at 05:19:17 UTC from IEEE Xplore. Restrictions apply.

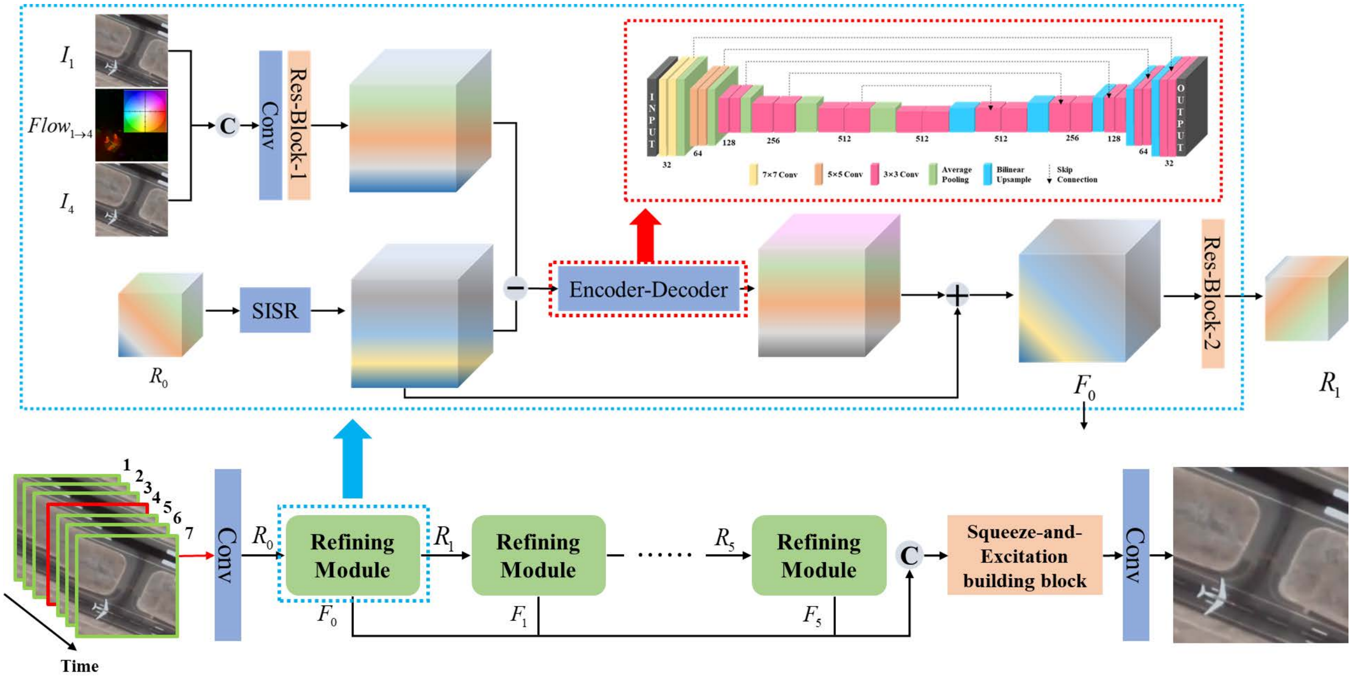

Fig. 1. The overall network structure of the proposed method

that the network can learn adaptively. In each refinement, the detailed features are enhanced through a deep Encoder-Decoder structure. Each reference frame participates in a refinement once and gets an output, ultimately aggregating the outputs of each refinement to produce our final SR target frame. A temporal attention mechanism is introduced to modulate each refinement's output, and the experiment shows our method is superior to EDVR.

2. METHODOLOGY

2.1. Overview of network structure

The overall structure of the proposed network is shown in Fig. 1. It can be seen that its core module is the refining module. Our network uses 7 consecutive frames as the input. We define the intermediate frame as the target frame (marked by a red box) and a total of 6 frames on the left and right as the reference frames (marked by a green box). Each reference frame participates in a refinement process, so a total of 6 refinement outputs $F_i\left(i=0,1,\cdots,5\right)$ are obtained. Finally, $F_i\left(i=0,1,\cdots,5\right)$ are combined to generate the SR frame corresponding to the target frame.

For the convenience of explanation, we assume 7 consecutive frames $I_i\left(i=1,\cdots,7\right)$ as input, and I_4 is the target frame. Suppose that the input LR frame size is $h\times w\times c$, where $c=3$ represents the number of RGB channels. The SR scale is set to r. The first refinement process is carried out as follows: firstly, PyFlow [5] is used to calculate the optical flow $Flow_{1\to4}$ between reference frame I_1 and the target frame I_4. Note that the number of optical flow channels is 2. Then, they are concatenated on the channel dimension to obtain the input of size $h\times w\times 8$, and a 3×3 convolution is followed to obtain the LR features of size $h\times w\times 256$. And then, after the first residual block with deconvolution, we obtain the HR features with the size of $\left(h\times r\right)\times\left(w\times r\right)\times64$. The LR frame with size $h\times w\times3$ is convolved with a 3×3 convolution to obtain the LR features R_0 of size $h\times w\times64$, then it will pass a SISR method named DBPN [6] which is widely used in recurrent structure to obtain HR features of size $\left(h\times r\right)\times\left(w\times r\right)\times64$. After that, we subtract this HR features from the HR features learned from reference frame, and the residual features which represent complementary information are enhanced through an UNet structure shown in Fig. 1. The enhanced residual features are added to the HR features obtained by DBPN. In this way, the network can be forced to learn the details that are missing from the HR features obtained by the SISR method. Then, we get the first refinement output F_0. Finally, in preparation for the next refinement, the HR features are downsampled to the LR features through the second residual block.

2.2. Deep enhancement of residual features

The learned residual features contain the missing information in the results of the SISR method, which is the key to the VSR. Actually, the residual features mainly represent the image texture, edge, and other high-frequency

Authorized licensed use limited to: Wuhan University. Downloaded on February 26,2022 at 05:19:17 UTC from IEEE Xplore. Restrictions apply.

Table 1. The following table shows the average PSNR and SSIM results for all frames of each test video

	Bicubic	**EDVR**	**Ours**		**Bicubic**	**EDVR**	**Ours**
000	35.996/0.9540	42.483/0.9862	**43.558/0.9881**	**004**	36.999/0.9501	40.732/0.9735	**40.813/0.9742**
001	32.278/0.9430	38.430/0.9790	**39.106/0.9801**	**005**	37.439/0.9565	41.688/0.9795	**41.694/0.9796**
002	36.502/0.9494	44.008/0.9878	**44.519/0.9887**	**006**	35.826/0.9502	40.609/**0.9782**	**40.644**/0.9781
003	38.551/0.9715	43.678/0.9850	**44.009/0.9855**	**007**	33.401/0.9265	37.616/0.9656	**37.748/0.9662**

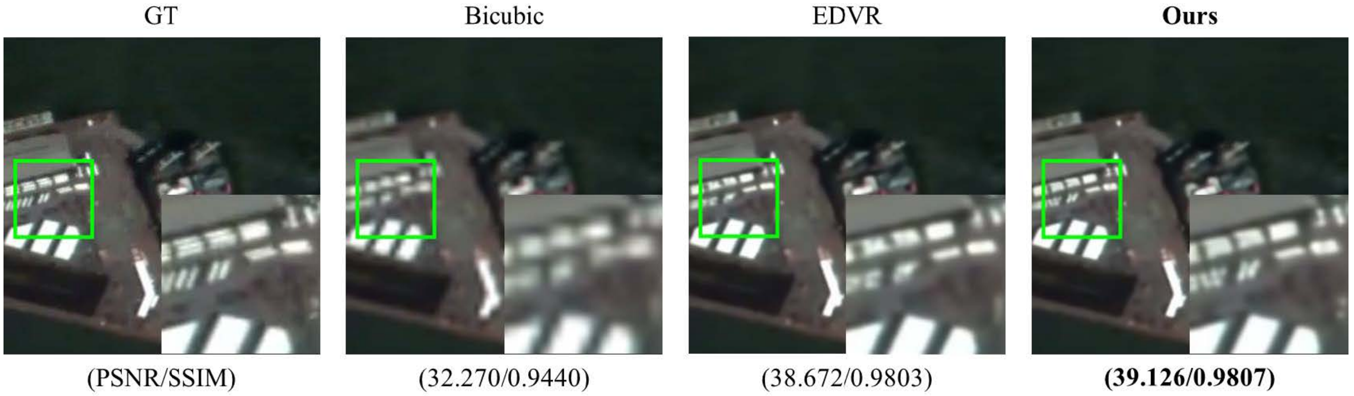

Fig. 2. ×4 SR results on test set 001 'wharf' scene. For better comparison, we select a frame from the test video 001 and partially zoom the region marked by green box to show more details.

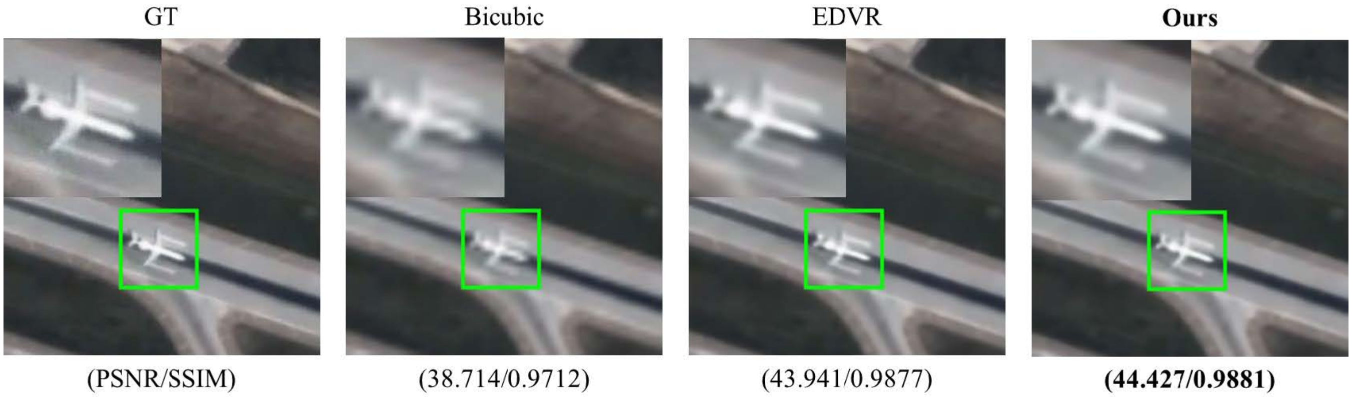

Fig. 3. ×4 SR results on test set 004 'plane1' scene.

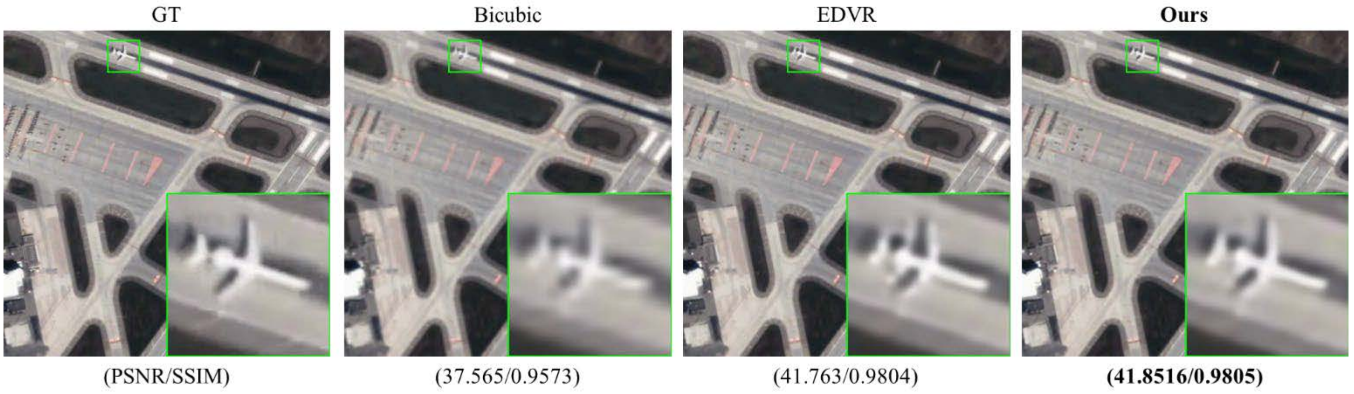

Fig. 4. ×4 SR results on test set 007 'plane2' scene.

information. Such detailed information tends to be sparse. If the network layers are too deep, detailed features will be lost in the deep features along with the convolution operation. However, if the network layers are too shallow, it is likely that the features will not be deep enough to be fully expressed.

Authorized licensed use limited to: Wuhan University. Downloaded on February 26,2022 at 05:19:17 UTC from IEEE Xplore. Restrictions apply.

To balance this relationship, we chose an UNet structure which is deep enough to learn the deep residual features. As shown in Fig. 1, in the classical UNet structure, the features are first continuously downsampled and then upsampled to restore the original size. In this way, the structural characteristics of the features can be well guaranteed. Simultaneously, to ensure that the sparse detailed features are not lost in the deeper part of the network, we add a skip connection between the shallow layer and the deep layer, so the shallow features can also be well transmitted to the deep layer. Such an Unet structure can realize the deep expression of features while maintaining the detailed information well.

2.3. Temporal attention mechanism

Generally speaking, the reference frame closer to the target frame has a higher structural similarity with the target frame. But the reference frame farther away from the target frame may also contain rich detailed information. Obviously, the complementary information they provide is different so the network needs to learn the difference automatically. Simply concatenating features on channel dimension is equal to treating each channel fairly, which does not consider the different contributions of different temporal distance features to the recovery of target frames. After concatenating $F_i\left(i=0,1,\cdots,5\right)$ on the channel dimension, the temporal attention is equivalent to the channel attention. So we chose the same plug-and-play channel attention module as in [7] for better integration.

3. EXPERIMENTAL RESULTS

Training Set. We have 10 Jilin-1 videos, as a result of the last two videos have a large reflective area. We only use the first 8 of them to build our training set. The original video frame size is 4096×2160. We crop each video into a video clip with a size of 640×640 in a non-overlapping way. Each video clip contains 100 consecutive frames, and a total of 192 video clips are obtained to build the training set.

Test Set. The test set contains 8 video clips. In the first 5 videos, we crop 5 scenes respectively as our test set (000-004). Each test video contains 100 consecutive frames with the frame size of 256×256. In addition, we crop 3 scenes (005-007) in the last two videos, each of them contains 200 consecutive frames with the frame size of 640×640.

Training details. Patch size is set to 64×64 and the minibatch size is set to 2. Adam with momentum to 0.9 is used as the optimizer. The initial learning rate is $10e-5$ and decays by a factor of 10 for half of total 30 epochs. The deep learning framework is Pytorch1.2 and we take 3 days to train our model on 2 NVIDIA RTX 2080 Ti GPU.

Results. We focus on ×4 SR. In the simulation experiment, PSNR and SSIM are used as our quantitative evaluation indexes. For the 8 test clips, we take the average PSNR and SSIM of all frames in each clip as the final result seen in Table. 1. It can be seen that our method achieves the best PSNR and SSIM results in all test sets. (The SSIM result on the test set 006 is only 0.0001 lower than the EDVR)

In terms of qualitative results, we show the ×4 SR result of a certain frame and display its local details. EDVR and our method both restore more texture and detail information than the Bicubic method. In the scene of 'wharf' in Fig. 2, by noting the goods distributed in strips on the ground, it can be observed that our model can distinguish them well, while the result of EDVR model cannot. In the scene of 'plane1' shown in Fig. 3 and the scene of 'plane2' in Fig. 4, our model appears more sharpened on the edge of the wing of the plane and has a better recovery effect on the tail of the plane. In addition, EDVR produces obvious artifacts near the tail of the plane, and our model is very good at restoring the actual shape.

4. CONCLUSION

In this paper, a recurrent refinement network for satellite video data SR is proposed. Through recurrent refinement, complementary information enhanced by an UNet structure is continuously learned from reference frames and supplemented to SISR result to finally aggregate the SR frame corresponding to the target frame. The experiment on the video of Jilin-1 illustrates our model achieves the best results both quantitatively and qualitatively.

5. REFERENCES

[1] Yue L, Shen H, Li J, *et al*. “Image super-resolution: The techniques, applications, and future,” *Signal Processing*, 2016, 128: 389-408.

[2] Caballero J, Ledig C, Aitken A, *et al*. “Real-time video super-resolution with spatio-temporal networks and motion compensation,” *IEEE Conference on Computer Vision and Pattern Recognition(CVPR)*. 2017: 4778-4787.

[3] Xue T, Chen B, Wu J, *et al*. “Video enhancement with task-oriented flow,” *International Journal of Computer Vision*, 2019, 127(8): 1106-1125.

[4] Wang X, Chan K C K, Yu K, *et al*. EDVR: “Video restoration with enhanced deformable convolutional networks,” *IEEE Conference on Computer Vision and Pattern Recognition Workshops*. 2019: 1-10.

[5] Liu C. “Beyond pixels: exploring new representations and applications for motion analysis,” *Massachusetts Institute of Technology*, 2009.

[6] Haris M, Shakhnarovich G, Ukita N. *et al*. “Deep back-projection networks for super-resolution,” *IEEE conference on computer vision and pattern recognition(CVPR)*. 2018: 1664-1673.

[7] Hu J, Shen L, Sun G. *et al*. “Squeeze-and-excitation networks,” *IEEE conference on computer vision and pattern recognition(CVPR)*. 2018: 7132-7141.

Authorized licensed use limited to: Wuhan University. Downloaded on February 26,2022 at 05:19:17 UTC from IEEE Xplore. Restrictions apply.