SELF-LEARNED VIDEO SUPER-RESOLUTION WITH AUGMENTED SPATIAL AND TEMPORAL CONTEXT

Zejia Fan[†], *Jiaying Liu*[†], *Wenhan Yang*[†], *Wei Xiang*[‡], *Zongming Guo*[†⋆]

[†]Wangxuan Institute of Computer Technology, Peking University, Beijing, China
[‡]Bigo, Beijing, China

ABSTRACT

Video super-resolution methods typically rely on paired training data, in which the low-resolution frames are usually synthetically generated under predetermined degradation conditions (*e.g.*, Bicubic downsampling). However, in real applications, it is labor-consuming and expensive to obtain this kind of training data, which limits the practical performance of these methods. To address the issue and get rid of the synthetic paired data, in this paper, we make exploration in utilizing the internal self-similarity redundancy within the video to build a Self-Learned Video Super-Resolution (SLVSR) method, which only needs to be trained on the input testing video itself. We employ a series of data augmentation strategies to make full use of the spatial and temporal context of the target video clips. The idea is applied to two branches of mainstream SR methods: frame fusion and frame recurrence methods. Since the former takes advantage of the short-term temporal consistency and the latter of the long-term one, our method can satisfy different practical situations. The experimental results show the superiority of our proposed method, especially in addressing the video super-resolution problems in real applications.

***Index Terms*—** Video Super-Resolution, Self-Learning, Augmentation, Spatial Context, Temporal Context

1. INTRODUCTION

Super-Resolution (SR) aims at turning Low-Resolution (LR) frames/images into the corresponding High-Resolution (HR) ones. Since video transmission is widely applied, there is a growing demand for video quality improvement, especially for Video Super-Resolution (VSR). Comparing with single-image SR [1–5] that only utilizes the spatial dependency, VSR models [6–12] can additionally utilize temporal dependency among frames, which derive more promising results. Based on the way to process each input frame, the VSR methods can be categorized into two classes: *frame fusion* and *frame recurrence*. The *frame fusion* VSR models take several consecutive frames as input and work in a *many-to-one* way. Usually, the motion between adjacent frames and the central frame is first estimated, followed by motion compensation to align the adjacent frames to the central frame. Then, VSR models carry out SR reconstruction to predict the central frame with these aligned frames. In comparison, the *frame recurrence* model builds an *one-to-one sequence-to-sequence* mapping which transforms the current input frame into the HR prediction progressively with support of information aggregated from the previous sequence process. This category is capable to maintain long-term memory information of the whole sequence for better effectiveness and efficiency.

Most VSR models are trained with full supervision, *i.e.*, large amounts of paired LR-HR data. This might come across the domain gap issue when applied to real scenarios. Since a large paired LR-HR dataset in real scenes is difficult to collect, supervised VSR models usually use synthetic down-sampled *ideal* data, which is generated by Bicubic down-scaling kernel with anti-aliasing. However, shooting conditions in real scenes might vary, *e.g.*, camera shaking, equipment parameters, and air conditions, which leads to a domain gap between the synthetic and real as shown in Fig. 1(b).

To address these issues, recently more efforts are put into the direction of unsupervised/self-supervised SR. Some [13, 14] carry out unsupervised learning on unpaired LR-HR datasets while others [15, 16] propose to train on the input LR image and exploit the image-specific information. However, fewer works focus on the self-supervised VSR problem.

In this paper, we propose a Self-Learned Video Super-Resolution (SLVSR) framework, which takes a single input video sequence for model training and is plug-and-play for most of the VSR deep neural networks. The framework does not rely on the training phase on external datasets and can be implemented with modest computation resources. The main idea is shown in Fig. 1(a). The proposed SLVSR generates training pairs from the single video input according to the patch recurrence in natural images and videos. Benefiting from the diversified data augmentation, our network can learn

⋆ Corresponding author.

This work is supported by the National Key Research and Development Program of China under Grant No. 2018AAA0102702, the National Natural Science Foundation of China under Contract No.62172020, State Key Laboratory of Media Convergence Production Technology and Systems, and Key Laboratory of Science, Technology and Standard in Press Industry (Key Laboratory of Intelligent Press Media Technology).

to perform high-quality super-resolution reconstruction with limited training data by exploiting both spatial and temporal context. Experiments show that SLVSR framework produces desirable results on the blind blur kernel case and the real-world video case with both high-quality details and well-preserved temporal-domain continuity.

2. SELF-LEARNED VIDEO SUPER-RESOLUTION

2.1. Motivation

In [17, 18], it is commonly observed that, *the small patches are recurring across scales within a single natural image*. Namely, the visual entropy inside a single image is much smaller than in a general external collection of images. This discovery provides the theoretical support, the possibility and the potential superiority to implement the self-supervised image/video super-resolution. The natural images/videos themselves contain abundant information. Since the recurrence happens across different scales, using the target natural image alone can produce enough reference patches to support SISR, which has been demonstrated in [15].

2.2. Framework

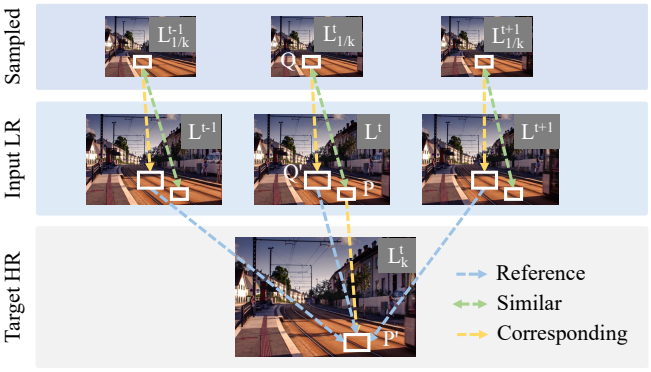
Based on the above consideration, we introduce the proposed Self-Learned Video Super-Resolution (SLVSR) framework. Fig. 1(a) demonstrates the our main idea. Given a low-resolution frame sequence $S = \{L^1, \cdots, L^n\}$, where n is the number of frames, and $S_k = \{L_k^1, \cdots, L_k^n\}$ represents the sequence up-sampled from S with a scale factor of k. With an input LR sequence S, SLVSR first generates the training examples from the testing frame itself by down-scaling each frame in the LR sequence S to get $S_{1/k} = \{L_{1/k}^1, \cdots, L_{1/k}^n\}$. Each frame in S can be coupled with the corresponding frame in $S_{1/k}$ to generate the training examples.

When we perform SR on the frame L^t, we first find a similar patch pair P and Q in L^t and $L_{1/k}^t$, respectively. Then, the internal spatial and temporal redundancies can be used to obtain useful clues for SR:

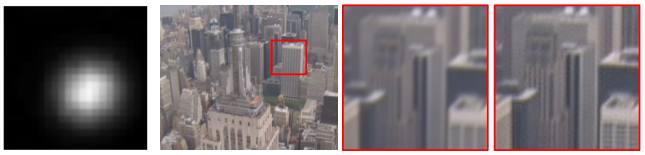
- **Cross-Scale Spatial Redundancy**: Let P' be the parent patch of P in L_k^t, Q' be the parent patch of Q in L^t, the LR-HR correlation between Q and Q' can offer useful guidance for the reconstruction of P' from P, which can be learned by our model.
- **Temporal Redundancy**: The neighboring frames of L^t can be aligned to L^t, and the patches relative to Q and Q' in these neighboring frames can form the constraints on the HR result.

2.3. Spatial and Temporal Context Utilization

In the self-supervised learning process, the training samples come from the input videos and the down-sampling versions



(a) Self-Learned Video Super-Resolution (SLVSR). Sampled images are down-sampled from Input LR images. The definition of the marks is shown in Sec. 2.2.



(b) Performance when the blind kernel is unknown.

Fig. 1. An overview of the work. (a) The key idea of the proposed self-learned video super-resolution. With an input LR video, SLVSR generates training examples from the testing input frame itself. (b) Supervised methods obtain unsatisfactory performance when the blur kernel is unknown, while our method can well handle the case with zero-shot learning.

of them along both the spatial and temporal dimensions. If the scale factor is k, in the training process, the scale proportion among the input, output, and the final target is $1 : k : k^2$. Compared with the original input, the motion amplitude of the objects in the down-sampled input video is reduced by k. Sampling videos with a larger temporal dilation leads to a motion amplitude more similar to that in the objects of the original LR video frames, which is more beneficial to model the motion of the input frames and the related frame alignment. Intuitively, the fast movement of an object becomes slower after the down-sampling in the spatial domain and can provide more useful guidance on modeling the normal object motion with similar structures in the input LR frames.

As for temporal context augmentation, we mainly apply a video sampling strategy, including *sampling with dilated interval among* $\{1, \cdots, k\}$ and *sampling inversely* in the time domain. There are two classes of deep VSR models distinguished by how temporal information is used:

- **Frame Fusion.** Frame fusion methods work in a *many-to-one* fashion, mapping successive LR frames into one SR output frame. In order to exploit the temporal context, we sample frame batches with different dilation factors. Consider L_k as a center frame from

input sequence, $\{L^{t-2i}, L^{t-i}, L^t, L^{t+i}, L^{t+2i}\}$ as a common five-frame batch with a dilation factor of i, $i \in \{1, \cdots, k\} \cup \{-k, \cdots, -1\}$. The downsampled batch and the center frame L_k will be added to the training dataset. From batches with different dilation factors, the VSR network can acquire more diverse information about motion modeling and LR-HR mapping.

- **Frame Recurrence.** Frame recurrence methods perform SR in a *one-to-one sequence-to-sequence* fashion to generate the SR result with the current LR image and the information of existing SR frames. We also sample the frames with different temporal dilation factors from the input sequence. Given input sequence $\{L^1, L^2, \cdots, L^{t-1}, L^t\}$, we first add sequences like $\{L^i, L^{2i}, \cdots, L^{t-i}, L^t\}$, $i \in \{1, \cdots, k\}$ and the versions in the inverse order into the training dataset as the HR sequence and then generate training sequence pairs from the augmented dataset.

2.4. Network Architecture

Our SLVSR framework can be applied to most VSR methods and brings improvement in performance. In our paper, we by default adopt VESPCN [19], a light-weighted frame fusion SR method, as the VSR model baseline. It estimates the displacement among neighboring frames and the center frame, then aligns these neighboring frames with a fast multi-resolution spatial transformer network based on CNN. Several adjacent frames after alignment are stacked together as low-resolution 3D data and 3D convolutions are used to perform the SR reconstruction. Similar procedures are performed when applying SLVSR to the frame recurrence method, such as FRVSR [9].

2.5. Training Data Generation

We generate k times down-sampled sequence $S_{1/k}$ from S. As pointed out in Sec. 1, the conventionally set *ideal* downsampling method brings in distortions in real scenarios. To address this problem, we generalize the single-image SR method KernelGAN [20] to the video input case. KernelGAN estimates the down-sampling kernel through internal learning. More specifically, the method trains a generator, which maps HR input to LR space as blur kernel, and a discriminator to judge whether a patch comes from a generator or the original image. In this way, Generative Adversarial Network (GAN) estimates a blur kernel that accurately generates the LR space similar to the input image patch distribution. As for the single video input, we treat all patches in the video constitute the input patch space and randomly sample frames in the spatial and temporal domain for both generator and discriminator. The trained blur kernel is shared among the video frames. In this way, we enrich the training data with more diverse blur kernels while improving computation efficiency.

Table 1. **Quantitative comparison** on Vid4 for 2× SR. Red and blue indicate the best and the second best performance, respectively. Evaluation carries out on the Y (luminance) channel with PSNR/SSIM metrics.

Method	RCAN (1 Frame)	DUF (7 Frames)	ZSSR (1 Frame)
Calendar	21.46/0.6787	20.08/0.6139	21.54/0.6808
City	25.41/0.6293	23.84/0.5784	25.24/0.6261
Foliage	24.13/0.6453	22.54/0.5697	24.11/0.6431
Walk	25.32/0.7794	23.65/0.7527	25.31/0.7855
Average	24.08/0.6832	22.53/0.6287	24.05/0.6839
Method	KernelGAN (1 Frame)	SinGAN (1 Frame)	SLVSR (3 Frames)
Calendar	22.17/0.7705	20.71/0.6118	25.10/0.8431
City	26.07/0.7026	24.74/0.5744	26.93/0.7459
Foliage	25.76/0.7835	24.15/0.6021	28.66/0.8651
Walk	26.96/0.8433	24.43/0.7465	27.86/0.8574
Average	25.24/0.7750	23.45/0.6337	27.14/0.8279

2.6. Implementation Details

As for data augmentation, we applies four rotations (0°, 90°, 180°, 270°), reflection and affine transformation to expand the dataset. As for the training process, we use $L1$ loss and ADAM [21] optimizer. The initial learning rate is set to $2 * 10^{-4}$. We periodically take a linear fit of the reconstruction error of generated training examples to adjust the learning rate. If the standard deviation is greater by a factor than the slope of the linear fit and the absolute value of the slope is relatively small, we divide the learning rate by 10. The training process ends with a learning rate smaller than 10^{-8}.

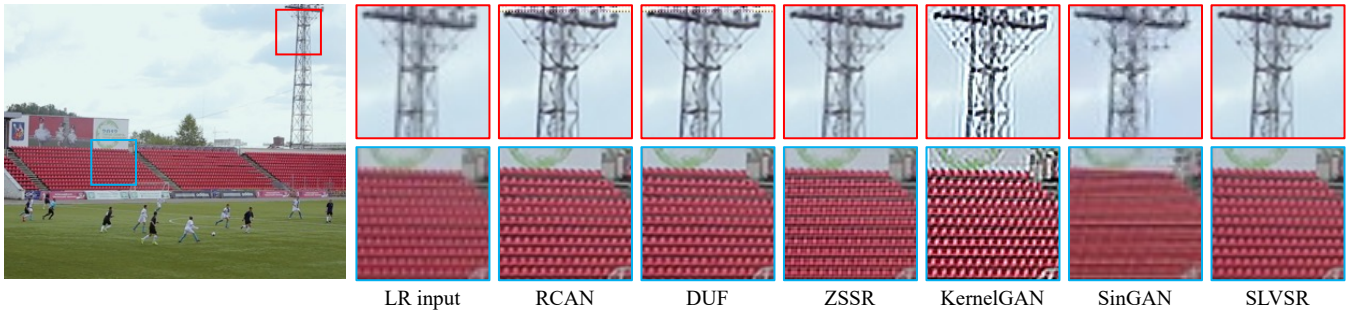
3. EXPERIMENTS

Experimental Setup. We test on Vid4 [22], which is widely used as a standard benchmark for VSR, and contains four sequences: city, calendar, walk and foliage. Although the performance on this test set is commonly reported in the literature, the motion between frames is relatively small and simple in these video sequences. To make up for this deficiency, we randomly pick out 20 video sequences from the Vimeo-90k [23] and name it the Vimeo-20. All the video clips are downloaded from the website vimeo.com and each is a 7-frame sequence, with a fixed resolution of 448×256. The data is also used as the low-resolution real-world clips. The compared methods include RCAN [24], DUF [6], ZSSR [15], and KernelGAN [20]. More experimental results are provided online[1].

Blind Blur Kernel Case. We adopt an-isotropic Gaussian kernels to generate LR frames for evaluation. We randomly sample the blur kernel for each video clip to generate the LR video frames from the HR ones and then measure the performance of each method with quantitative metrics. We show the quantitative comparison with Table 1. It is observed that our SLVSR achieves considerably better results than other

[1]https://zahrafan.github.io/icassp22_SLVSR/

(a) Evaluation on videos generated by blind blur kernel.



(b) Evaluation on the real-world video.

Fig. 2. **Visualization comparison** among different methods.

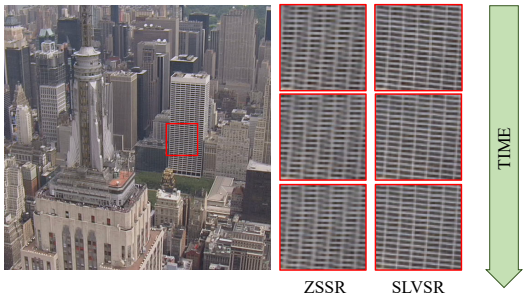


Fig. 3. **Temporal domain continuity.** We evaluate the performance in three continuous frames. SLVSR is evidently affected less by aliasing.

methods, with at least 0.9dB PSNR gains. The qualitative results are demonstrated in Fig. 2(a). From the results, we can observe that, most methods fail to restore the clear results, *e.g.* RCAN, ZSSR. KernelGAN generates obvious visual artifacts. Comparatively, SLVSR provides clear visual results with fewer artifacts.

Real-World Videos. Our efforts on the VSR task ultimately aim at applying this technology to real-world videos. Therefore, the performance of the model on real-world low-quality input is very worthy of our attention. Fig. 2(b) shows that our SLVSR can produce satisfactory results on real videos. Compared with ZSSR, SLVSR is able to reconstruct clearer boundaries.

Temporal Domain Continuity. As shown in Fig. 3, when looking continuously along with the temporal domain, SLVSR produces results with better continuity. Moreover, our results have more distinct edges, and the edges of different objects do not tend to mix together.

Table 2. **Ablation study on temporal augmentation.** The performance is measured by the PSNR and SSIM on Vid4 testing set on Y (luminance) channel.

Method	Temporal Aug.	PSNR/SSIM
SLVSR+VESPCN		32.13/0.9370
SLVSR+VESPCN	✓	32.25/0.9409
SLVSR+FRVSR		30.91/0.9199
SLVSR+FRVSR	✓	30.99/0.9216

Ablation Study on Data Augmentation. To explore the effect of data augmentation, we perform ablation studies. The results can be seen in Table 2. We use SLVSR+VESPCN to represent the frame fusion method and SLVSR+FRVSR for frame recurrence method. With the temporal data augmentation, SLVSR+VESPCN obtains a 0.15dB gain in PSNR and SLVSR+FRVSR obtains a 0.08dB gain in PSNR. These results show the effectiveness of temporal data augmentation.

4. CONCLUSION

In this work, we propose a plug-and-play self-learned video super-resolution framework, which exploits the temporal and spatial context of the input video sequence, relying on neither external examples nor prior training. The framework can be widely used on a variety of VSR networks. Experimental results demonstrate the superior performance and the generalization of our designs. Considering the rich information in the external data, it will be our future work to explore a way to properly combine the advantage of self-learning and supervised learning on external datasets.

5. REFERENCES

[1] C. Dong, C. C. Loy, K. He, and X. Tang, "Image super-resolution using deep convolutional networks," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 38, no. 2, pp. 295–307, 2016.

[2] B. Lim, S. Son, H. Kim, S. Nah, and K. M. Lee, "Enhanced deep residual networks for single image super-resolution," in *Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition Workshops*, July 2017.

[3] J. Kim, J. K. Lee, and K. M. Lee, "Accurate image super-resolution using very deep convolutional networks," in *Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition*, 2016.

[4] T. Dai, J. Cai, Y. Zhang, S. Xia, and L. Zhang, "Second-order attention network for single image super-resolution," in *Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition*, 2019.

[5] Y. Zhang, Y. Tian, Y. Kong, B. Zhong, and Y. Fu, "Residual dense network for image super-resolution," in *Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition*, 2018.

[6] Y. Jo, Seoung W. Oh, J. Kang, and S. J. Kim, "Deep video super-resolution network using dynamic upsampling filters without explicit motion compensation," in *Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition*, 2018.

[7] Y. Tian, Y. Zhang, Y. Fu, and C. Xu, "TDAN: Temporally-deformable alignment network for video super-resolution," in *Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition*, June 2020.

[8] X. Wang, K. C. Chan, K. Yu, C. Dong, and C. C. Loy, "EDVR: Video restoration with enhanced deformable convolutional networks," in *Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition Workshops*, 2019.

[9] M. S. Sajjadi, R. Vemulapalli, and M. Brown, "Frame-recurrent video super-resolution," in *Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition*, 2018.

[10] J. Lin, Y. Huang, and L. Wang, "FDAN: Flow-guided deformable alignment network for video super-resolution," *arXiv preprint arXiv:2105.05640*, 2021.

[11] K. C. Chan, X. Wang, K. Yu, C. Dong, and C. C. Loy, "BasicVSR: The search for essential components in video super-resolution and beyond," in *Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition*, 2021.

[12] S. Y. Kim, J. Oh, and M. Kim, "FISR: deep joint frame interpolation and super-resolution with a multi-scale temporal loss," in *Proc. AAAI Conf. on Artificial Intelligence*, 2020, vol. 34.

[13] Y. Yuan, S. Liu, J. Zhang, Y. Zhang, C. Dong, and L. Lin, "Unsupervised image super-resolution using cycle-in-cycle generative adversarial networks," in *Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition Workshops*, 2018.

[14] A. Lugmayr, M. Danelljan, and R. Timofte, "Unsupervised learning for real-world super-resolution," in *Proc. IEEE Int'l Conf. Computer Vision Workshop*, 2019.

[15] A. Shocher, N. Cohen, and M. Irani, "Zero-shot super-resolution using deep internal learning," in *Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition*, 2018.

[16] T. R. Shaham, T. Dekel, and T. Michaeli, "SinGAN: Learning a generative model from a single natural image," in *Proc. IEEE Int'l Conf. Computer Vision*, 2019.

[17] D. Glasner, S. Bagon, and M. Irani, "Super-resolution from a single image," in *Proc. IEEE Int'l Conf. Computer Vision*, 2009.

[18] M. Zontak and M. Irani, "Internal statistics of a single natural image," in *Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition*, 2011.

[19] J. Caballero, C. Ledig, A. Aitken, A. Acosta, J. Totz, Z. Wang, and W. Shi, "Real-time video super-resolution with spatio-temporal networks and motion compensation," in *Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition*, 2017.

[20] S. Bell-Kligler, A. Shocher, and M. Irani, "Blind super-resolution kernel estimation using an internal-GAN," in *Proc. Annual Conf. Neural Information Processing Systems*, 2019.

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

[22] C. Liu and D. Sun, "A bayesian approach to adaptive video super resolution," in *Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition*, 2011.

[23] T. Xue, B. Chen, J. Wu, D. Wei, and W. T. Freeman, "Video enhancement with task-oriented flow," *Int'l Journal of Computer Vision*, vol. 127, no. 8, pp. 1106–1125, 2019.

[24] Y. Zhang, K. Li, K. Li, Lichen. Wang, B. Zhong, and Y. Fu, "Image super-resolution using very deep residual channel attention networks," in *Proc. IEEE European Conf. Computer Vision*, 2018.