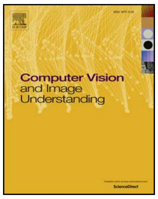
Contents lists available at ScienceDirect

Computer Vision and Image Understanding

journal homepage: www.elsevier.com/locate/cviu



# Video super-resolution based on spatial-temporal recurrent residual networks

Wenhan Yang[a], Jiashi Feng[b], Guosen Xie[c], Jiaying Liu[*,1,a], Zongming Guo[a], Shuicheng Yan[d]

[a] Institute of Computer Science and Technology, Peking University, Beijing 100871, PR China
[b] Department of Electrical and Computer Engineering, National University of Singapore, 117583, Singapore
[c] NLPR, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, PR China
[d] Artificial Intelligence Institute, Qihoo 360 Technology Company, Ltd., Beijing 100015, PR China

## ARTICLE INFO

*Keywords:*
Spatial residue
Temporal residue
Video super-resolution
Inter-frame motion context
Intra-frame redundancy

## ABSTRACT

In this paper, we propose a new video Super-Resolution (SR) method by jointly modeling intra-frame redundancy and inter-frame motion context in a unified deep network. Different from conventional methods, the proposed Spatial-Temporal Recurrent Residual Network (STR-ResNet) investigates both spatial and temporal residues, which are represented by the difference between a high resolution (HR) frame and its corresponding low resolution (LR) frame and the difference between adjacent HR frames, respectively. This spatial-temporal residual learning model is then utilized to connect the intra-frame and inter-frame redundancies within video sequences in a recurrent convolutional network and to predict HR temporal residues in the penultimate layer as guidance to benefit estimating the spatial residue for video SR. Extensive experiments have demonstrated that the proposed STR-ResNet is able to efficiently reconstruct videos with diversified contents and complex motions, which outperforms the existing video SR approaches and offers new state-of-the-art performances on benchmark datasets.

## 1. Introduction

Video super-resolution (SR) aims to produce high-resolution (HR) video frames from a sequence of low-resolution (LR) inputs. In recent years, video super-resolution has been drawing increasing interest from both academia and industry. Although various HR video devices have been developed constantly, it is still highly expensive to produce, store and transmit HR videos. Thus, there is a great demand for modern SR techniques to generate HR videos from LR ones.

The video SR problem, as well as other signal super-resolution problems, can be summarized as restoring the original scene $\mathbf{x}_t$ from its several quality-degraded observations $\{\mathbf{y}_t\}$. Typically, the observation can be modeled as

$$\mathbf{y}_t = \mathbf{D}_t\mathbf{x}_t + \mathbf{v}_t, \; t = 1, \; \dots, T. \tag{1}$$

Here $\mathbf{D}_t$ encapsulates various signal quality degradation factors at the time instance $t$, *e.g.*, motion blur, defocus blur and down-sampling. Additive noise during observation at that time is denoted as $\mathbf{v}_t$. Generally, the SR problem, i.e., solving out $\mathbf{x}_t$ in Eq. (1), is an ill-posed linear inverse problem that is rather challenging. Thus, accurately estimating $\mathbf{x}_t$ demands either sufficient observations $\mathbf{y}_t$ or proper priors on $\mathbf{x}_t$.

All video SR methods can be divided into two classes: reconstruction-based and learning-based. Reconstructed-based methods (Baker and Kanade, 1999; Farsiu et al., 2004; He and Kondi, 2006; Kanaev and Miller, 2013; Liu and Sun, 2014; Omer and Tanaka, 2009; Rudin et al., 1992) craft a video SR process to solve the inverse estimation problem of (1). They usually perform motion compensation at first, then perform deblurring by estimating blur functions in $\mathbf{D}_t$ of (1), and finally recover details by local correspondences. The hand-crafted video SR process cannot be applicable for every practical scenario of different properties and perform not well to some unexpected cases.

In contrast, learning-based methods handle the ill-posed inverse estimation by learning useful priors for video SR from a large collection of videos. Typical methods include recently developed deep learning-based video SR methods (Huang et al., 2015; 2017; Liao et al., 2015a) and give some examples of non-deep learning approaches. In Liao et al. (2015a), a funnel shape convolutional neural network (CNN) was developed to predict HR frames from LR frames that are aligned by optical flow in advance. It shows superior performance on recovering HR video

[*] Corresponding author.
*E-mail address:* liujiaying@pku.edu.cn (J. Liu).

[1] This work was supported by National Natural Science Foundation of China under contract no. 61772043, with additional support by the State Scholarship Fund from the China Scholarship Council.

https://doi.org/10.1016/j.cviu.2017.09.002
Received 24 January 2017; Received in revised form 25 June 2017; Accepted 16 September 2017

frames captured in still scenes. However, this CNN model suffers from high computational cost (as it relies on time-consuming regularized optical flow methods) as well as visual artifacts caused by complex motions in the video frames. In Huang et al. (2015); 2017), a bidirectional recurrent convolutional network (BRCN) was employed to model the temporal correlation among multiple frames and further boost the performance for video SR over previous methods.

However, previous learning-based video SR methods that learn to predict HR frames directly based on LR frames, suffer from following limitations. First, these methods concentrate on exploiting between-frame correlations and does not *jointly* consider the intra- and inter-frame correlations that are both critical for the quality of video SR. This unfavorably limits the capacity of the network for recovering HR frames with complex contents. Second, the successive input LR frames are usually highly correlated with the whole signal of the HR frames, but are not correlated with the high frequency details of these HR images. In the case where dominant training frames present slow motion, the learned priors hardly capture hard cases, such as large movements and shot changes, where neighboring frames distinguished-contributed operations are needed. Third, it is desirable for the joint estimation of video SR to impose priors on missing high frequency signals. However, in previous methods, the potential constraints are directly enforced on the estimated HR frames.

To solve the above-mentioned issues, in this work, we propose a unified deep neural network architecture to *jointly* model the intra-frame and the inter-frame correlation in an end-to-end trainable manner. Compared with previous (deep) video SR methods (Huang et al., 2015; 2017; Liao et al., 2015a), our proposed deep network model does not require explicit computation of optical flow or motion compensation. In addition, our proposed model unifies the convolutional neural networks (CNNs) and recurrent neural networks (RNNs) which are known to be powerful in modeling sequential data. Combining the spatial convolutional and temporal recurrent architectures enables our model to capture spatial and temporal correlations jointly. Specially, it models spatial and temporal correlations among multiple video frames jointly. The temporal residues of HR frames are predicted based on input LR frames along with their temporal residues to further regularize estimation of the spatial residues.

This architectural choice enables the network to handle the videos containing complex motions in a moving scene, offering pleasant video SR results with few artifacts in a time-efficient way.

More concretely, we propose a **S**patial **T**emporal **R**ecurrent **Res**idual **Net**work (STR-ResNet) for video SR as show in Fig. 1. As aforementioned, SRT-ResNet models spatial and temporal correlations among multiple video frames jointly. In STR-ResNet, one basic component is the spatial residual CNN (SRes-CNN) for single frame SR, which has a bypass connection for learning the residue between LR and HR feature maps. SRes-CNN is able to capture the correlation information among pixels within a single frame, and tries to recover an HR frame based on its corresponding LR frame through utilizing such correlations. Then, STR-ResNet stacks multiple SRes-CNNs together with recurrent connections between them. The global recurrent architecture captures the temporal contextual correlation and recovers the HR frame using both its corresponding LR frame and its adjacent frames. To better model inter-frame motions, STR-ResNet takes not only multiple LR frames but also the residue of these adjacent LR frames as inputs and tries to predict the temporal residues of HR frames in the penultimate layer. An HR frame is thus recovered by STR-ResNet by summing up its corresponding LR frame and the predicted spatial residue via the SRes-CNN component, under the guidance of the predicted temporal residue from adjacent frames via recurrent residual learning.

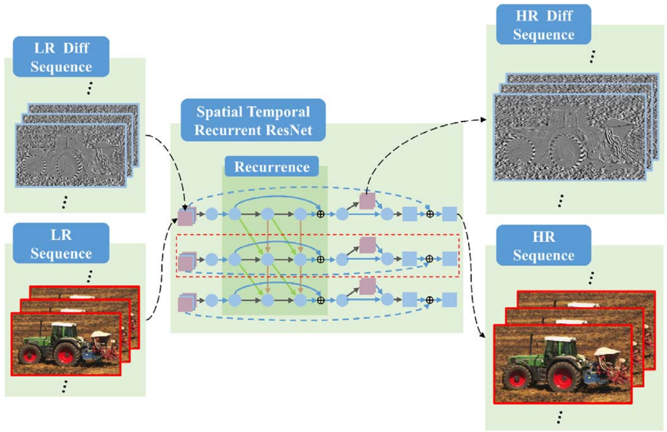


**Fig. 1.** The architecture of our proposed spatial-temporal recurrent residual network (STR-ResNet) for video SR. It takes not only the LR frames but also the differences of these adjacent LR frames as the input. Some reconstructed features are constrained to predict the differences of adjacent HR frames in the penultimate layer.

By separating the video frames into LR observations and the spatial residue within a single frame, the low frequency parts of HR frames and LR frames are untangled. Thus, the models can only focus on describing high-frequency details. By considering the temporal residues, in both their prediction path from LR temporal residues to HR temporal residues and their connection to spatial residues, the proposed STR-ResNet models both the spatial and temporal correlations jointly and achieves outstanding video SR performance with relatively low computational complexity.

In summary, we make the following contributions in this work to solving the challenging video SR problem:

- We propose a novel deep convolutional neural network architecture specifically for video SR. It follows a joint spatial-temporal residual learning and aims to predict the HR temporal residues which further facilitate the predictions of spatial residues and HR frames. By embedding the temporal residue prediction, the proposed architecture is capable of implicitly modeling the motion context among multiple video frames for video SR. It provides high-quality video SR results on benchmark datasets with relatively low computational complexity.
- To the best of our knowledge, the proposed STR-ResNet is the first research attempt to incorporate the bypass connection in a deep network to embed the joint spatial-temporal residue prediction and model temporal correlations in video frame sequences for video processing. The incorporated residual architecture implicitly models inter-frame motion context and is demonstrated to be beneficial for video SR.
- We are also among the first to investigate and unify the spatial convolutional, temporal recurrent and residual architectures into a single deep neural network to solve video SR problems. Extensive experiments on video SR benchmark datasets clearly demonstrate the contribution of each component to the overall performance.

The rest of this paper is organized as follows. Related work is briefly reviewed in Section 2. In Section 3, we introduce our spatial-temporal residual learning. Then, we construct a deep network to model it step-by-step and present the details of the proposed STR-ResNet, which models both spatial and temporal redundancies jointly in a unified network, as well as its constituent SRes-CNN in Section 4. Experimental results are presented in Section 5. More analysis and discussion on our method are provided in Section 6. Concluding remarks are given in Section 7.

## 2. Related work

Single image super-resolution was first investigated by Irani and Peleg (1991). By now, it can be divided into two categories: reconstruction-based and learning-based. Reconstruction-based methods adopt regularizations, such as gradient histogram (Sun et al., 2011),