Grouped Spatio-Temporal Alignment Network for Video Super-Resolution

Mingxuan Lu and Peng Zhang

Abstract—Video super-resolution based on CNNs has recently achieved significant progress. The existing popular super-resolution methods usually impose the optical flow between neighboring frames for temporal alignment. However, estimation of accurate optical flow is hard and expends greater computation. To address this problem, we propose a novel grouped spatio-temporal alignment network (GSTAN) that effectively incorporates spatio-temporal information in a hierarchical way. The input sequence is divided into several groups corresponding to different frame rates. These groups provide complementary information, which is helpful to restore missing textures for the reference frame. Specifically, each group employs deformable 3D convolution to incorporate spatio-temporal information, which avoids artifacts from explicit motion estimation. In addition, a Gated-Dconv information filter is proposed to control information flow to focus on the fine details. Finally, these groups provide complementary information, which is integrated with the inter-group fusion module. Extensive experiments have demonstrated our method achieves state-of-the-art SR performance on several benchmark datasets.

Index Terms—Video super-resolution, spatio-temporal alignment, inter-group fusion.

I. INTRODUCTION

The goal of video super-resolution (VSR) is to recover high-resolution (HR) video frames from the low-resolution (LR) ones. This technique has great value in many applications such as satellite imagery [1] and video surveillance [2]. Compared to single-image super-resolution [3], [4], [5], [6], video super-resolution poses an unsolved challenge of how to fully employ the spatio-temporal dependency since video sequence provides additional temporal information.

Previous VSR methods usually use the optical flow between a reference frame and neighboring frames to align consecutive frames. Therefore, the quality of optical flow estimation directly affects the restoration of video super-resolution. For example, Wang et al. [7] proposed an optical flow reconstruction network to learn temporal details between consecutive frames through optical flow estimation. Chan et al. [8] proposed a bidirectional propagation network based on optical flow alignment to aggregate information in the spatio-temporal dimension. These methods mainly use optical flow estimation for alignment, which leads to the introduction of artifacts and greatly affect the restoration.

To avoid such problems, [9], [10], [11] implicitly integrated motion information among low-resolution frames without computing optical flow. However, these methods ignore the complementary information between video frames and cannot fully employ the spatio-temporal information. In this work, we divide the input sequence consisting of $2N+1$ frames into $N$ groups based on decoupled motion. Instead of aligning $2N$ neighboring frames to the reference frame with optical flow directly, we exploit the complementary information between these groups to restore missing textures for the reference frame. With the powerful modeling capability of 3D deformable convolution [12], the spatio-temporal information of each group is jointly integrated. To further obtain fine information flow for efficient restoration, we propose a Gated-Dconv information filter to focus on fine details at the end of the spatio-temporal alignment module. Finally, the inter-group features providing different complementary information are integrated with temporal attention and deeply fused with a bottleneck layer including 2D convolution units to produce a high-resolution reference frame. Overall, our method follows a hierarchical way, which can adaptively integrate information from groups of different frame rates.

In summary, our contributions in this letter are as follows:

1) We propose a novel grouped spatio-temporal alignment network (GSTAN), which can make use of complementary information of frame-rate-aware groups and fully exploit the spatio-temporal information using deformable 3D convolution.

2) We propose a Gated-Dconv information filter to control feature transformation. It suppresses less informative features and transmits useful information for effective super-resolution.

3) Our network achieves excellent performance through the evaluation of commonly used datasets with high computational efficiency.

II. METHOD

A. Frame Feature Extraction

Given $2N+1$ consecutive low-resolution video frames consisting of one reference frame $I^L_t$ and $2N$ neighboring frames $\{I^L_{t-N}, I^L_{t-1}, I^L_{t+1}, \ldots, I^L_{t+N}\}$. Our goal is to reconstruct HR reference frame $I^H_t$ from the reference frame $I^L_t$ with the help of neighboring frames. In this work, the overall network
IEEE SIGNAL PROCESSING LETTERS, VOL. 29, 2022

Fig. 1. Architecture of the proposed GSTAN consists of three parts: frame feature extraction, spatio-temporal alignment, and inter-group fusion. Here, we show the framework with seven video frames.

B. Spatio-Temporal Alignment

In contrast to previous work [9], [11], [14], the feature maps are used directly for spatio-temporal alignment. Considering the contributions of spatio-temporal alignment with different temporal distances are not equal. We divide the feature maps \( \{F_D^i\}_{i=1}^7 \) into 3 groups \( \{F_D^D_n, F_D^D_{i+n}, F_D^D_{i+2n}\}_{n=1}^3 \) at different frame rates. For these groups, we integrate spatio-temporal information with a residual deformable 3D convolution (ResD3D) block. It has long been demonstrated that deformable 3D convolution has a large spatial receptive field [15]. Therefore, it models spatio-temporal information efficiently. In our work, we use five ResD3D blocks to model group-wise features \( F_g \), as shown in Fig. 3(a).

To improve representation learning, we propose a Gated-Dconv information filter at the end of the spatio-temporal alignment module, as shown in Fig. 3(b). It consists of two parts: 1) gating mechanism, and 2) depthwise convolutions. The gating mechanism can be regarded as the product of two parallel transmitted information, one of which is activated with the GELU non-linearity [16]. To learn local video structure for an effective restoration, we use depth-wise convolution to encode information from spatially neighboring pixel positions. Given a tensor \( X \in \mathbb{R}^{H \times W \times C} \), the process is formulated as:

\[
\hat{X} = W_p \text{Gating}(X) + X \quad (1)
\]

\[
\text{Gating}(X) = \phi \left( W_{pd}^1 (\text{LN}(X)) \right) \odot W_{pd}^2 (\text{LN}(X)) \quad (2)
\]

where \( W_{pd} \) represents the \( 1 \times 1 \) point-wise convolution and the \( 3 \times 3 \) depth-wise convolution, \( \odot \) denotes element-wise multiplication, \( \text{LN} \) is the layer normalization [17] and \( \phi \) represents the GELU non-linearity. Overall, the module controls information flow to focus on the fine details complementary to other groups. A discussion of this Gated-Dconv information filter will be given in the ablation study.

C. Inter-group Fusion

To integrate features from these groups, we proposed an inter-group fusion module with temporal attention [18]. The previous section illustrates that the input sequence is divided into these groups, which usually contain complementary
information. Based on the facts that 1) neighboring frames with slow frame rates are more similar to the reference one. They have richer information for SR recovery, and 2) neighboring frames with fast frame rates also have particular information that the reference one is missing. Hence, this module serves as guidance to integrate features from these groups.

At the end of the spatio-temporal alignment module, the group-wise feature $F^g_n$ channel dimension is 64. The corresponding single channel feature map $F^c_n$ is computed applying a ReLU and a 3 × 3 convolution layer. They are further concatenated and each position $(x, y)$, across channels is calculated for the attention maps $M(x, y)$ with softmax function. Then, attention weighted feature $\tilde{F}^g_n$ is calculated by element-wise multiplication of $M_n(x, y)$ and $F^g_n$.

$$M_n(x, y) = \frac{e^{F^c_n(x, y)}}{\sum_{i=1}^{N} e^{F^c_i(x, y)}}$$  \hspace{1cm} (3)

The inter-group fusion module aims to integrate the complementary information of these groups and produce a high-resolution version of the residual map. We first concatenate those features $\tilde{F}^g_n$ to aggregate them along the temporal axis. A bottleneck layer including 2D convolution units is employed to reduce channels and compress feature dimensions. Finally, like most super-resolution methods, the fused feature is upsampled with a sub-pixel layer for SR reconstruction.

III. EXPERIMENTS

A. Implementation Details

Similar to [11], [19], [20], we employ Vimeo-90K [21] containing 64612 training samples with spatial resolution 448 × 256 as the training set. The corresponding LR sequences with spatial resolution 112 × 64 are obtained by bicubic downsampling with 4 times. We augment the training data by randomly rotating and flipping [22]. To evaluate our model, we employ the Vimeo-90K-T [21], Vid4 [23], and SPMC [23] set for performance evaluation. We use PSNR and SSIM as quantitative metrics, which are widely used for the task of video super-resolution.

We implement all experiments in Pytorch with an Nvidia RTX 3090 GPU. The networks are optimized by an ADAM optimizer [24] ($\beta_1 = 0.9$, $\beta_2 = 0.999$) with a batch size of 32 and supervised by Charbonnier loss ($\varepsilon$ is set to $10^{-3}$). The initial learning rate is set to $4 \times 10^{-4}$ and halved for every 20 epochs. We train the network for 200 epochs.

B. Ablation Study

1) Input Sequence Length: To demonstrate the importance of input sequence length, we fix the batch size to 32. The results of our network fed to input sequence with different lengths (3, 5 and 7) are shown in Fig. 4. The comparison shows that training with longer sequences is preferable (e.g., $L = 7$ has a clear and smooth pattern). Because training with longer sequence allows the network to employ long-term information more effectively. When the input sequence is 7, it is divided into three groups, which can be recovered by using the complementary information from the other two groups. The input sequence is 3 or 5, the less complementary information is employed. However, with a fixed computational overhead, the increased input sequence length when training model inevitably leads to a decrease in batch size. Training with a small batch size causes more unstable gradients.

2) Intra-Group Spatio-Temporal Alignment: To verify the performance improvement introduced by Gated-Dconv information filter, we compare the performance change after adding the module. In addition, the results with different numbers of residual D3D blocks are shown in Fig. 5. Finally, it can be observed that the PSNR and SSIM values of the network adding this module are increased by 0.09 and 0.003 on average. It achieves the best performance with 5 residual D3D. That illustrates the module controls information flow to focus on the fine details, which improves representation learning.

3) Inter-Group Feature Fusion: We experiment with different ways of organizing the input sequence. First, we feed the input sequence directly into the network. In addition, we experiment with other ways of grouping: $\{142, 345, 647\}$ and $\{123, 345, 567\}$. As shown in Table I, No-grouping

With a constraint of $B \times L = 224$, we adjust $L$ to 9 (input sequence length) and $B$ (batch size) to 24. As shown in Fig. 4 that restoration becomes worse, which indicates employing a longer sequence is preferable when a computational constraint is imposed. Therefore, we train the network with a batch size is 32 and the input sequence length is 7.
method has the worst performance, which illustrates integrating temporal information hierarchically is a more effective way. \{123, 345, 567\} indicates that each group does not have a reference frame. Its performance is inferior to the other two groups of methods. Compared to \{142, 345, 647\} including arbitrary neighboring frames, our grouping method \{345, 246, 147\} improves 0.04 dB, which is attributed to the effectiveness of frame rate-based grouping in employing temporal information.

### C. Comparisons With State-of-the-Art Methods

In this section, we compare our method with several state-of-the-art SR approaches, including a single image SR method (RCAN) [4] and 5 video SR methods (i.e., VSRnet [25], VESPCN [26], SOF-VSR [7], TDANet [11], and D3Dnet [12]). For a fair comparison, the public implementations are adopted and we train all methods on the Vimeo-90k set.

The $4 \times$ quantitative reconstruction results are listed in Table II. The test results of various data sets show that our method achieves the highest evaluation scores among all compared methods. Compared with the suboptimal approach, our method improves the average index by 0.33/0.011. Fig. 6 shows qualitative results achieved by different methods. It can be observed that our method can recover finer details (e.g., the sharp edge of building window lines). That is, our method fully incorporates spatio-temporal information in a hierarchical way. We also provide a comparison of computational efficiency listed in Table III. As compared with RCAN, our GSTAN achieves improvements in performance and computational efficiency. The FLOPs of GSTAN are 74.3% of optimal explicit alignment-based BasicVSR [8] and 22.7% of VRT [27] based on the prevalent model Transformer. Furthermore, the inference speed of our method is a bit faster than BasicVSR and is about four times faster than VRT.

### IV. Conclusion

In this work, we proposed a grouped spatio-temporal alignment network (GSTAN) for video SR. Our network exploits spatio-temporal information in a novel way. The proposed method maintains temporal consistency and also reconstructs high-quality HR frames. Extensive experiments on several benchmark datasets achieve state-of-the-art SR performance. In the future, we will further explore addressing general real-world degraded videos for practical application with our approach.
REFERENCES

[1] M. Deudon et al., “Highres-net: Recursive fusion for multi-frame super-resolution of satellite imagery,” 2020, arXiv:2002.06460.

[2] Y. Xiao, X. Su, Q. Yuan, D. Liu, H. Shen, and L. Zhang, “Satellite video super-resolution via multiscale deformable convolutional alignment and temporal grouping projection,” IEEE Trans. Geosci. Remote Sens., vol. 60, pp. 1–19, 2022.

[3] C. Dong, C. C. Loy, K. He, and X. Tang, “Image super-resolution using deep convolutional networks,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 38, no. 2, pp. 295–307, Feb. 2016.

[4] Y. Zhang, Y. Tian, Y. Kong, B. Zhong, and Y. Fu, “Residual dense network for image super-resolution,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., 2020, pp. 3360–3369.

[5] X. Xiang, Y. Tian, Y. Zhang, Y. Fu, J. P. Allebach, and C. Xu, “Zooming LU AND ZHANG: GROUPED SPATIO-TEMPORAL ALIGNMENT NETWORK FOR VIDEO SUPER-RESOLUTION 2197

[6] Y. Chen, S. Liu, and X. Wang, “Learning continuous image representation with local implicit image function,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., 2021, pp. 8628–8638.

[7] 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/CVF Conf. Comput. Vis. Pattern Recognit., 2021, pp. 4947–4956.

[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/CVF Conf. Comput. Vis. Pattern Recognit. Workshops, 2019, pp. 1594–1603.

[9] X. Xiang, Y. Tian, Y. Zhang, Y. Fu, J. P. Allebach, and C. Xu, “Zooming slow-mo: Fast and accurate one-stage space-time video super-resolution,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., 2020, pp. 3370–3379.

[10] X. Ying, L. Wang, Y. Wang, W. Sheng, W. An, and Y. Guo, “Deformable 3D convolution for video super-resolution,” IEEE Signal Process. Lett., vol. 27, pp. 1500–1504, 2020.

[11] J. Liang, J. Cao, G. Sun, K. Zhang, L. Van Gool, and R. Timofte, “SwinIR: Image restoration using swin transformer,” in Proc. IEEE/CVF Int. Conf. Comput. Vis., 2021, pp. 1833–1844.

[12] J. Zang, L. Wang, Z. Liu, Q. Zhang, G. Hua, and N. Zheng, “Attention-based temporal weighted convolutional neural network for action recognition,” in Proc. IFIP Int. Conf. Artif. Intell. Appl. Innov., Springer, 2018, pp. 97–108.

[13] J. Zang, L. Wang, Z. Liu, Q. Zhang, G. Hua, and N. Zheng, “Attention-based temporal weighted convolutional neural network for action recognition,” in Proc. IFIP Int. Conf. Artif. Intell. Appl. Innov., Springer, 2018, pp. 97–108.

[14] J. Liang, J. Cao, G. Sun, K. Zhang, L. Van Gool, and R. Timofte, “SwinIR: Image restoration using swin transformer,” in Proc. IEEE/CVF Int. Conf. Comput. Vis., 2021, pp. 3360–3369.

[15] J. Liang, J. Cao, G. Sun, K. Zhang, L. Van Gool, and R. Timofte, “SwinIR: Image restoration using swin transformer,” in Proc. IEEE/CVF Int. Conf. Comput. Vis., 2021, pp. 3360–3369.