Video Super-Resolution Model Based on Group Feedback Fusion Mechanism (Postprint)

Authors: Zhang Qingwu, Chi Xiaoyu, Zhu Jian, Chen Bingfeng, Ruichu Cai

Date: 2022-05-18T00:00:00+00:00

Abstract

Video super-resolution (VSR) aims to generate a high-resolution version of a reference frame by utilizing information from multiple adjacent frames. Numerous existing VSR studies have concentrated on effectively aligning neighboring frames to facilitate better fusion of adjacent frame information, yet seldom have they investigated the important step of adjacent frame information fusion itself. To address this problem, we propose a video super-resolution model based on a group feedback fusion mechanism (GFFMVSR). Specifically, after neighboring frame alignment, the aligned video sequence is input into a first-stage temporal attention module. The sequence is then partitioned into several groups, with each group sequentially passing through an intra-group fusion module to achieve preliminary fusion. Subsequently, the fusion results from different groups are processed by a second-stage temporal attention module. Each group is then fed into the feedback fusion module sequentially, where a feedback mechanism is utilized to fuse information across different groups. Finally, the fused result is output for reconstruction. Empirical verification demonstrates that this model exhibits strong information fusion capabilities and surpasses existing models in both objective evaluation metrics and subjective visual effects.

Full Text

Preamble

Vol. 39 No. 10
Application Research of Computers (ChinaXiv Cooperative Journal)
Accepted Paper

Video Super-Resolution Model Based on Group Feedback Fusion Mechanism

Zhang Qingwu[1], Chi Xiaoyu[2], Zhu Jian[1], Chen Bingfeng[1]†, Cai Ruichu[1]
(1. School of Computers, Guangdong University of Technology, Guangzhou

510006, China;
2. Qingdao Research Institute, Beihang University, Qingdao 266000, China)

Abstract: Video super-resolution (VSR) aims to generate a high-resolution version of a reference frame by exploiting information from multiple adjacent frames. While many existing VSR works have focused on how to effectively align adjacent frames to better fuse their information, few have investigated the crucial step of adjacent frame information fusion itself. To address this gap, we propose a video super-resolution model based on group feedback fusion mechanism (GFFMVSR). Specifically, after aligning adjacent frames, the aligned video sequence is fed into the first temporal attention module. The sequence is then divided into several groups, each of which undergoes preliminary fusion through a group-wise intra-group fusion module. Subsequently, the fusion results from different groups pass through a second temporal attention module. Each group is then sequentially fed into a feedback fusion module, which leverages a feedback mechanism to fuse information across different groups. Finally, the fused result is output for reconstruction. Experiments demonstrate that this model possesses strong information fusion capabilities and outperforms existing models in both objective evaluation metrics and subjective visual quality.

Keywords: video super-resolution; temporal attention; feedback mechanism; group fusion

0 Introduction

Super-resolution (SR) refers to the process of reconstructing a high-resolution (HR) image from its corresponding low-resolution (LR) counterpart. Depending on the number of input frames, SR tasks can be categorized into two types: single-image super-resolution (SISR) and video super-resolution (VSR). This paper focuses on the VSR task, which has attracted widespread attention in computer vision and image processing research due to its broad application prospects —for instance, when surveillance footage needs to be enlarged for person or license plate identification, or when videos are projected onto high-definition displays for visual enhancement.

The first deep learning-based SISR algorithm, SRCNN, was proposed by Dong et al. [**?**], consisting of three convolutional layers that learn an end-to-end nonlinear mapping from LR to HR images, demonstrating impressive potential. Since then, numerous deep learning methods have been applied to SISR. For example, VDSR [**?**], inspired by VGG [**?**], employs a deeper convolutional network architecture. Li et al. [**?**] proposed SRFBN, a network architecture that uses feedback connections to leverage more contextual information for correcting low-level feature learning. Pan et al. [**?**] introduced FFAMSR, a network architecture applying residual-in-residual (RIR) structures and combining spatial and coordinate attention for thorough feature extraction and reuse. Despite achieving state-of-the-art performance, these networks suffer from high computational costs and

memory footprint, limiting their deployment on mobile devices. To address this, lightweight networks such as FALSR-A [?] and SMSR [?] have been proposed.

In the VSR domain, Huang et al. [?] proposed BRCN, a bidirectional recurrent convolutional network that models long-term temporal information across multiple frames to improve VSR quality. Caballero et al. [?] introduced VESCPN, which jointly trains optical flow estimation and spatio-temporal networks in an end-to-end manner for efficient VSR. Tao et al. [?] proposed SPMC, which simultaneously achieves motion compensation and upsampling through a designed sub-pixel motion compensation module. Inspired by the inherent spatio-temporal learning capability of 3DCNN, Kim et al. [?] proposed 3DSRNet, which stacks multiple 3D convolutional layers for VSR while avoiding direct motion alignment. Jo et al. [?] introduced DUF, which utilizes 3DCNN to mine spatio-temporal information and predicts a dynamic upsampling filter [?] for implicit motion compensation and upsampling, thereby replacing pixel-level optical flow estimation and alignment. Haris et al. [?] proposed RBPN, which leverages spatial and temporal information through recurrent encoder-decoder modules. TDAN [?] and EDVR [?] applied deformable convolution to VSR and proposed a temporally deformable alignment module that achieves motion alignment at the feature level.

Hupé [?] and Gilbert [?] et al. discovered that in human cognitive theory, feedback connections connecting cortical visual areas can transmit reflected signals from higher-order regions to lower-order regions for utilization. Building upon this, Zamir et al. [?] proposed a feedback mechanism network applicable to computer vision. In recent years, this mechanism has been incorporated into network architectures for various visual tasks [?, ?, ?], demonstrating promising results. To the best of our knowledge, however, feedback mechanisms have not yet been applied to VSR research. Inspired by prior work [?, ?], we hypothesize that since feedback mechanisms [?] allow networks to carry historical information to influence the learning of new inputs, could fused results containing partial adjacent frame information similarly influence the fusion of remaining adjacent frames? To this end, we propose a video super-resolution model based on group feedback fusion mechanism (GFFMVSR). The principle of our feedback scheme is that results containing partial adjacent frame information can facilitate better fusion of the remaining adjacent frame information.

The main contributions are:
a) We propose a video super-resolution model based on grouping and feedback mechanism principles, which can effectively fuse high-level information from aligned frames and improve video reconstruction capability.
b) We introduce a group feedback mechanism into the VSR domain, providing a novel adjacent frame information fusion method to enhance spatio-temporal information fusion performance.
c) We incorporate dual temporal attention into the model, where temporal attention modules can capture important information hidden in adjacent frames, enabling the network to recover clearer video frames with richer details.

1.1 Network Architecture

As shown in Figure 1 [Figure 1: see original paper], the video super-resolution model based on group feedback fusion mechanism consists of five main components: Feature Extraction and Alignment Module (FEAM), Intra-Group Fusion Module (IGFM), Dual Temporal Attention Module (DTAM), Feedback Fusion Module (FFM), and Rebuild Module (RM). Blue arrows indicate feedback fusion, while green arrows denote global residual skip connections. The network' s task is to reconstruct a high-resolution version of the reference frame from an input video sequence of N frames. We define the input video sequence as $\{I_{r-N}^{LR}, \dots, I_r^{LR}, \dots, I_{r+N}^{LR}\}$, the super-resolved version of the reference frame as I_r^{SR}, the ground-truth high-resolution version as I_r^{HR}, deconvolution operation as $Deconv(s, n)$, and convolution operation as $Conv(s, n)$, where s is the filter size and n is the number of filters.

The FEAM extracts and aligns features from adjacent frames, as shown in Equation (1):

$$\{F_{r-N}^{a}, \dots, F_r^{a}, \dots, F_{r+N}^{a}\} = f_{FEAM}(\{I_{r-N}^{LR}, \dots, I_r^{LR}, \dots, I_{r+N}^{LR}\})$$

where $f_{FEAM}(\cdot)$ represents the feature extraction and alignment operation. In FEAM, feature extraction is simply implemented through downsampling with strided convolution operations, while alignment follows the multi-scale deformable convolution-based method proposed in EDVR [**?**] (i.e., the PCD alignment module). Readers are referred to EDVR [**?**] for detailed information about the PCD alignment module.

The aligned adjacent frames are then fed into Temporal Attention Module 1 (TAM_1) to compute similarity between adjacent frames and the reference frame, which facilitates intra-group information fusion. This operation is shown in Equation (2):

$$\{F_{r-N}^{a'}, \dots, F_r^{a'}, \dots, F_{r+N}^{a'}\} = f_{TAM_1}(\{F_{r-N}^{a}, \dots, F_r^{a}, \dots, F_{r+N}^{a}\})$$

where $f_{TAM_1}(\cdot)$ represents the operation of Temporal Attention Module 1.

The sequence $\{F_{r-N}^{a'}, \dots, F_r^{a'}, \dots, F_{r+N}^{a'}\}$ is divided into N groups, each representing a specific frame rate. Each group sequence is fed into a parameter-shared IGFM to achieve preliminary fusion within the group, producing the fused feature sequence defined as $\{G_1, G_2, \dots, G_N\}$ (the IGFM module will be detailed in Section 1.2).

To extract information useful for subsequent reconstruction, a second temporal attention module (TAM_2) with the same structure as TAM_1 is inserted after IGFM, forming the Dual Temporal Attention Module (DTAM). This temporal attention module will be elaborated in Section 1.3. The feature sequence after

TAM_2 is defined as $\{G'_1, G'_2, \ldots, G'_N\}$, with the operation shown in Equation (3):

$$\{G'_1, G'_2, \ldots, G'_N\} = f_{TAM_2}(\{G_1, G_2, \ldots, G_N\})$$

where $f_{TAM_2}(\cdot)$ represents the operation of Temporal Attention Module 2.

Cross-group information is further integrated through the feedback fusion module based on a feedback mechanism. As shown in Figure 2 [Figure 2: see original paper], the red dashed box in Figure 1 can be unfolded into T iterations ($t \in [1, T]$), representing one iteration in Figure 1 . To enable the hidden state in FFM to carry the concept of output, we connect the loss at each iteration. The loss function will be detailed in Section 1.5. The elements in sequence $\{G'_1, G'_2, \ldots, G'_N\}$ are fed into the FFM module sequentially to achieve feedback fusion. Additionally, G'_1 is treated as the initial hidden state F^0_{out}.

The t-th iteration of FFM receives the t-th group feature G'_t and the hidden state F^{t-1}_{out} from the previous iteration, producing the t-th output F^t_{out} of FFM. This operation is shown in Equation (4):

$$F^t_{out} = f_{FFM}(G'_t, F^{t-1}_{out})$$

where $f_{FFM}(\cdot)$ represents the FFM operation, and the actual feedback process is illustrated in Figure 2. The FFM module will be detailed in Section 1.4.

Since different groups contain different information after fusion, to emphasize important information, the feedback fusion results are fed into the reconstruction module to generate a residual image. As shown in Figure 1, the reconstruction module uses HR features and performs the operation shown in Equation (5):

$$Re_t = f_{RM}(F^t_{out})$$

where $f_{RM}(\cdot)$ represents the reconstruction module operation. The reconstruction module upscales the fused LR features to generate the network' s residual image.

Temporal attention can focus more effectively on features beneficial for subsequent reconstruction rather than treating all features equally. DTAM consists of two identical temporal attention modules with the structure shown in Figure 4 [Figure 4: see original paper], named TAM_1 and TAM_2. They focus on capturing temporal information and computing weights for feature sequences before and after group fusion, respectively, thereby improving information fusion effectiveness.

Finally, the high-resolution version of the reference frame I^{SR}_r is generated by adding the network-produced residual map to the bicubic upsampling of the input reference frame. This operation is shown in Equation (6):

$$I^{SR}_{r,t} = f_{up}(I^{LR}_r) + Re_t$$

where $f_{up}(\cdot)$ represents the upsampling kernel operation. The choice of upsampling kernel is arbitrary; here we use a bicubic upsampling kernel. After T

iterations, we obtain T SR versions of the reference frame $\{I_{r,1}^{SR}, I_{r,2}^{SR}, \dots, I_{r,T}^{SR}\}$. Notably, as the number of iterations increases, the reconstructed reference frame carries more adjacent frame information and becomes closer to the ground-truth HR version. Therefore, we select the result from the last iteration as the final reconstruction.

1.2 Intra-Group Fusion Module (IGFM)

Adjacent frames at greater temporal distances may contain less useful information. To fully utilize useful information, eliminate excessive irrelevant features, and improve subsequent feedback efficiency, preliminary non-feedback intra-group fusion is required before feedback fusion. The aligned feature sequence $\{F_{r-N}^{a'}, \dots, F_{r}^{a'}, \dots, F_{r+N}^{a'}\}$ is divided into groups. Unlike previous work, adjacent N frames are divided into N groups based on their temporal distance to the reference frame. The original sequence is rearranged as $\{G_1, G_2, \dots, G_N\}$, where each G_n consists of a subsequence containing the previous frame $F_{r-n}^{a'}$, the reference frame $F_{r}^{a'}$, and the next frame $F_{r+n}^{a'}$. It should be noted that the reference frame appears in every group (the specific reason is discussed in Section 2.2). The contributions of adjacent frames at different temporal distances are not equal, and grouping enables efficient information extraction and fusion of adjacent frames at different temporal distances under the guidance of the reference frame. Notably, our method can be easily generalized to arbitrary numbers of input frames.

For each group, the intra-group fusion module is deployed for feature fusion within the group. As shown in Figure 3 [Figure 3: see original paper], the front part of this module uses 3D convolutional layers with $3 \times 3 \times 3$ kernels to achieve spatio-temporal feature fusion for each group. Then, 15 2D units are applied in a 2D dense block to deeply integrate information within each group, producing the grouped feature sequence $\{G_1, G_2, \dots, G_N\}$. Each unit in the dense block sequentially consists of batch normalization [**?**] (BN), ReLU [**?**], 1×1 convolution, BN, ReLU, and 3×3 convolution. As done in [**?**], each 2D unit concatenates all previous feature maps as input. Finally, a 1×1 convolutional layer reduces the number of channels. The 2D dense block design is inspired by DUF [**?**]. To improve efficiency, the weights of the intra-group fusion module are shared across all groups, with effective modifications made to the data flow channels. The operation of this module is shown in Equation (7):

$$\{G_1, G_2, \dots, G_N\} = f_{IGFM}(\{G_1, G_2, \dots, G_N\})$$

where $f_{IGFM}(\cdot)$ represents the intra-group fusion module operation. Section 2.2 verifies the effectiveness of the proposed temporal grouping strategy.

1.3 Dual Temporal Attention Module (DTAM)

Inter-frame temporal relationships are crucial in VSR adjacent frame fusion (due to occlusion, blur regions, and parallax issues, different adjacent frames contain varying amounts of information). The goal of temporal attention is to compute similarity between feature sequences in an embedding space. Intuitively, in an embedding space, more attention should be paid to feature information that is more similar to the reference feature. In TAM_1, for each frame feature, the similarity distance (i.e., temporal attention map M_i^a) can be calculated as:

$$M_i^a = \text{sigmoid}(\theta(F_i^a) \cdot \phi(F_r^a))$$

where F_r^a is treated as the reference feature, and $\theta(\cdot)$ and $\phi(\cdot)$ are two embedding operations that can be implemented through simple convolutional filters. The sigmoid activation function restricts the output to $[0, 1]$, stabilizing gradient backpropagation. Note that the temporal attention map has the same size as the feature map F_i^a. The attention-weighted feature for each adjacent frame is computed as:

$$F_i^{a'} = F_i^a \odot M_i^a$$

where $\odot$ represents element-wise multiplication.

Similarly, for TAM_2, the reference feature is G_i'. Its temporal attention map M_i^g and attention-weighted features are computed as shown in Equations (10) and (11) (the same applies here). Notably, in TAM_2, the reference frame appears in every group, which is also the case in TAM_1.

1.4 Feedback Fusion Module (FFM)

The FFM module is shown in Figure 5 [Figure 5: see original paper]. The t-th iteration ($t \in [1, T]$) of FFM receives feedback information F_{out}^{t-1} to guide the fusion of the t-th group feature map G_t', then passes the more informative representation F_{out}^t to the next iteration and reconstruction module, forming a complete feedback process. To implement the feedback fusion function of FFM, the module sequentially contains three projection groups, where information flows effectively across levels through dense skip connections. Each projection group mainly includes upsampling and downsampling operations, which project HR features onto LR features to continuously refine the fused features. FFM is executed iteratively to effectively fuse the feature sequence $\{G_1', G_2', \ldots, G_N'\}$. The iterative process is unfolded as shown in Figure 2.

G_t' and F_{out}^{t-1} are concatenated and compressed to guide the fusion of input features G_t' through feedback information F_{out}^{t-1}, producing the input feature F_{in}^t for the refinement group:

$$F_{in}^t = C_1([G_t', F_{out}^{t-1}])$$

where $C_1(\cdot)$ represents the initial channel compression operation, and $[\cdot]$ denotes concatenation.

We define H_t^g and L_t^g as the HR and LR feature maps produced by the g-th projection group in FFM during the t-th iteration, where $g \in [1, 3]$. H_t^g can be obtained through:

$$H_t^g = \mathrm{Dec}_g([L_t^{g-1}, L_t^{g-2}, \ldots, L_t^1])$$

where $\mathrm{Dec}_g(\cdot)$ represents upsampling using $Deconv(k, m)$ at the g-th projection group. Similarly, L_t^g can be obtained by:

$$L_t^g = \mathrm{Conv}_g([H_t^g, H_t^{g-1}, \ldots, H_t^1])$$

where $\mathrm{Conv}_g(\cdot)$ represents downsampling using $Conv(k, m)$ at the g-th projection group. To reduce parameters and improve computational efficiency, we add channel compression operations before all $Conv(k, m)$ and $Deconv(k, m)$ operations except the first projection group.

To fully utilize useful information from each projection group, we perform feature fusion on the LR features produced by the projection groups (shown by red arrows in Figure 5) to generate the output of the FFM module:

$$F_{out}^t = C_3([F_{out}^{t-1}, L_t^1, L_t^2, L_t^3])$$

1.5 Loss Function

In this work, we choose L1 loss to optimize the proposed network. Although only the final result from the last iteration of the reconstruction sequence $\{I_{r,1}^{SR}, I_{r,2}^{SR}, \ldots, I_{r,T}^{SR}\}$ is used as the final output, during training we still need to connect intermediate results to the loss function to ensure that each FFM iteration maximally fuses useful information from the current input feature map. The loss function of the network can be expressed as:

$$Loss = \sum_{t=1}^{T} \| I_{r,t}^{SR} - I_r^{HR} \|_1$$

where θ represents the network parameters. W_t is a constant factor representing the contribution value of the SR result at each iteration. We set all W_t to 1, meaning each reconstructed SR result has equal contribution, enabling each iteration to fuse high-level information as much as possible.

2.1 Experimental Settings

Dataset. We adopt Vimeo-90k [**?**] as the training set, which is widely used for video super-resolution and contains approximately 90k 7-frame video clips.

We crop regions with a spatial resolution of 256×448 from the high-resolution video clips. Similar to [**?**, **?**], we generate 64×112 low-resolution video clips by applying a Gaussian blur kernel with standard deviation $\sigma = 1.6$ and $4\times$ downsampling. We evaluate our method on two popular benchmark datasets: Vid4 [**?**] and Vimeo90K-T [**?**]. Both benchmark datasets contain various motion and occlusion scenes, making them suitable for evaluating our method' s information fusion and high-resolution reconstruction capabilities.

Implementation Details. Unless otherwise specified, our network takes 7 video frames as input, i.e., $N = 3$. We use PReLU [**?**] as the activation function after all convolutional and deconvolutional layers except the final layer in each sub-network. We set $T = 6$, $k = 4$ with stride 4 and padding 2, and $m = 64$. We optimize using the Adam [**?**] optimizer with $\beta_1 = 0.9$ and $\beta_2 = 0.999$. No weight decay is used during training. The learning rate is initially set to 2×10^{-4} and then reduced by a factor of 0.5 every 8 epochs until training ends at 60 epochs. The mini-batch size is set to 2. Training data is augmented through flipping and rotation with a probability of 0.5.

All experiments are conducted on a GPU server equipped with Python 3.8, PyTorch 1.1, and an Nvidia 2080TI.

2.2 Ablation Studies

Grouping Experiments. We first organize the input sequence using different methods. A baseline method (denoted as Base1) simply stacks the input sequence along the temporal axis and feeds it into IGFM and FFM modules at once (without temporal attention modules in between). Here, the FFM module only executes once without feedback mechanism. In addition to our proposed grouping method $\{345, 246, 147\}$, we also tried other grouping strategies: $\{123, 345, 567\}$ and $\{345, 142, 647\}$. As shown in Table 1 , the grouping method $\{345, 246, 147\}$ achieves the highest PSNR, suggesting that including the reference frame in each group helps the model extract missing information from the reference frame. The suboptimal performance of $\{345, 142, 647\}$ can be attributed to the large information differences between adjacent frames at different temporal distances from the reference frame, which is detrimental to grouped information learning.

Component Ablation Experiments. To verify the role of each module, we use the grouping method $\{345, 246, 147\}$ mentioned above as the Base model (denoted as Base2). We then introduce Temporal Attention Module 1 (TAM_1), Temporal Attention Module 2 (TAM_2), and the feedback fusion mechanism (FFM) to the Base2 model respectively. Notably, disabling the feedback fusion mechanism after grouping is achieved by concatenating adjacent group features along the temporal dimension and executing the FFM module only once. Additionally, the complete model integrating TAM_1, TAM_2, and FFM is denoted

as GFFMVSR (Ours). With an upscaling factor of 4, the PSNR values on the Vid4 test set are shown in Table 2 .

As seen in rows 1, 2, 3, and 4 of Table 2, introducing TAM_1 and TAM_2 improves the Base2 model' s PSNR by 0.09 dB and 0.08 dB, respectively. When both temporal attention modules are introduced to form the dual temporal attention module, PSNR improves by 0.13 dB. As shown in rows 1 and 5, introducing the feedback fusion mechanism improves PSNR by 0.18 dB. The final row shows that the complete model with dual temporal attention and feedback fusion mechanism achieves the best performance, surpassing Base2 by 0.27 dB, confirming the validity of our proposed model.

2.3 Comparison with State-of-the-Art Methods

In this section, we compare our method with several state-of-the-art VSR methods, including TOFlow [**?**], DUF [**?**], RBPN [**?**], EDVR [**?**], MuCAN [**?**], Liu [**?**], PFNL [**?**], and VSR-Transformer [**?**]. TOFlow and RBPN both use optical flow for explicit motion estimation at the pixel level. EDVR adopts implicit motion estimation with stronger noise handling capability. DUF, MuCAN, and PFNL skip the motion estimation process. The last method specifically uses the latest Vision Transformer (ViT) [**?**] network for VSR tasks. We run publicly available code or carefully reproduce most methods ourselves, attempting to replicate the results reported in the original papers.

Vid4 Dataset. Table 3 shows quantitative results on Vid4, with data either computed by us or taken from the original papers. Y and RGB represent luminance and RGB channels, respectively, and "-" indicates unavailable values. As a downgraded version of GFFMVSR, GFFMVSR-S (which only uses Temporal Attention Module 1) achieves average PSNR/SSIM values of 27.43/0.8373 on the Y channel and 25.93/0.8186 on RGB channels, outperforming all other methods. With dual temporal attention, GFFMVSR-S becomes GFFMVSR, achieving even higher performance on both Y and RGB channels.

Qualitative results are shown in Figure 6 [Figure 6: see original paper]. We can see that GFFMVSR produces sharper edges and finer textures than other methods, validating the superiority of our approach. Additionally, to compare temporal consistency performance, we extract and visualize temporal profiles from the Calendar sequence in the Vid4 dataset (see Figure 7 [Figure 7: see original paper]). Temporal profiles are obtained by taking horizontal pixel rows at the same position across multiple consecutive frames (the red line in Figure 7) and stacking them vertically. We can observe that GFFMVSR produces the most consistent results with fewer flickering artifacts and more uniform line details compared to other methods.

Vimeo-90K-T Dataset. Vimeo-90K-T contains approximately 7k video clips selected from Vimeo-90K as a test set, covering numerous scenes with large

motion. Quantitative PSNR/SSIM results are shown in Table 4 , which also includes parameter counts for most methods. Our method far exceeds most state-of-the-art methods in PSNR and SSIM, such as TOFlow, DUF, RBPN, and MuCAN. The only exception is EDVR-L, whose model size is about four times larger than ours, and EDVR involves a pre-training process requiring substantial data and training time. Nevertheless, our method achieves comparable PSNR and slightly better SSIM.

Furthermore, qualitative results on Vimeo-90K-T are shown in Figure 8 [Figure 8: see original paper]. We can see that GFFMVSR can also produce visually convincing SR images on this challenging dataset.

---

3 Conclusion

This paper addresses the problem that feedback mechanisms existing in the human visual system have not been fully exploited in current video super-resolution models. We propose a novel end-to-end trainable video super-resolution network called GFFMVSR. By combining grouping principles with feedback mechanisms for VSR tasks, we effectively improve adjacent frame information fusion and target frame reconstruction quality. The input sequence is reorganized into several groups of subsequences with different frame rates, allowing hierarchical spatio-temporal information extraction, followed by intra-group fusion modules for preliminary group feature fusion. The feedback fusion mechanism efficiently learns and fuses newly input content through feedback information, mimicking human cognitive learning processes. Applying temporal attention at appropriate positions in the model to form the dual temporal attention model further encourages the model to focus on useful information fusion. Extensive experiments on several benchmark datasets demonstrate that our proposed model outperforms existing VSR methods both quantitatively and qualitatively.

References

[1] Dong Chao, Loy Chen Change, He Kaiming, et al. Learning a deep convolutional network for image super-resolution [C]// European Conference on Computer Vision. Springer, Cham, 2014: 184-199.
[2] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition [J]. arXiv preprint arXiv: 1409.1556, 2014.
[3] Kim J, Lee J K, Lee K M. Accurate image super-resolution using very deep convolutional networks [C]// Proceedings of The IEEE Conference on Computer Vision and Pattern Recognition. 2016: 1646-1654.
[4] Li Zhen, Yang Jinglei, Liu Zheng, et al. Feedback network for image super-resolution [C]// Proceedings of The IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019: 3867-3876.

[5] 盘展鸿, 朱鉴, 迟小羽, 等. 基于特征融合和注意力机制的图像超分辨率模型 [J]. 计算机应用研究, 2022, 39 (3): 5. (Pan Zhanhong, Zhu Jian, Chi Xiaoyu, et al. Image super-resolution model based on feature fusion and attention mechanism [J]. Application Research of Computers. 2022, 39 (3): 5.)
[6] Chu Xiangxiang, Zhang Bo, Ma Hailong, et al. Fast, accurate and lightweight super-resolution with neural architecture search [C]// 25th International Conference on Pattern Recognition (ICPR). IEEE, 2021:
[7] Wang Longguang, Dong Xiaoyu, Wang Yingqian. et al. Exploring sparsity in image super-resolution for efficient inference [C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021: 4917-4926.
[8] Huang Yan, Wang Wei, Wang Liang. Bidirectional recurrent convolutional networks for multi-frame super-resolution [J]. Advances in Neural Information Processing Systems, 2015, 28.
[9] Caballero J, Ledig C, Aitken A, et al. Real-time video super-resolution with spatio-temporal networks and motion compensation [C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017: 4778-4787.
[10] Tao Xin, Gao Hongyun, Liao Renjie, et al. Detail-revealing deep video super-resolution [C]// Proceedings of the IEEE International Conference on Computer Vision. 2017: 4472-4480.
[11] Kim S Y, Lim J, Na T, et al. 3dsrnet: Video super-resolution using 3d convolutional neural networks [J]. arXiv preprint arXiv: 1812.09079,
[12] Jo Y, Oh S W, Kang J, et al. Deep video super-resolution network using dynamic upsampling filters without explicit motion compensation [C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018: 3224-3232.
[13] Jia X, De Brabandere B, Tuytelaars T, et al. Dynamic filter networks for predicting unobserved views [C]// Proceedings ECCV 2016 workshops. 2016: 1-2.
[14] Haris M, Shakhnarovich G, Ukita N. Recurrent back-projection network for video super-resolution [C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019: 3897-
[15] Tian Yapeng, Zhang Yulun, Fu Yun, et al. Tdan: Temporally-deformable alignment network for video super-resolution [C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020: 3360-3369.
[16] Wang Xintao, Chan K C K, Yu Ke, et al. Edvr: Video restoration with enhanced deformable convolutional networks [C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. 2019: 0-0.
[17] Hupé J M, James A C, Payne B R, et al. Cortical feedback improves discrimination between figure and background by V1, V2 and V3 neurons [J]. Nature, 1998, 394 (6695): 784-787.
[18] Gilbert C D, Sigman M. Brain states: top-down influences in sensory processing [J]. Neuron, 2007, 54 (5): 677-696.

[19] Zamir A R, Wu T L, Sun L, et al. Feedback networks [C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017: 1308-1317.
[20] Carreira J, Agrawal P, Fragkiadaki K, et al. Human pose estimation with iterative error feedback [C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016: 4733-4742.
[21] Sam D B, Babu R V. Top-down feedback for crowd counting convolutional neural network [C]// Thirty-second AAAI Conference on Artificial Intelligence. 2018.
[22] Wang Xintao, Chan K C K, Yu Ke, et al. Edvr: Video restoration with enhanced deformable convolutional networks [C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. 2019: 0-0.
[23] Ioffe S, Szegedy C. Batch normalization: Accelerating deep network training by reducing internal covariate shift [C]// International Conference on Machine Learning. PMLR, 2015: 448-456.
[24] Glorot X, Bordes A, Bengio Y. Deep sparse rectifier neural networks [C]// Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics. JMLR Workshop and Conference Proceedings, 2011: 315-323.
[25] Huang Gao, Liu Zhuang, Van Der Maaten L, et al. Densely connected convolutional networks [C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017: 4700-4708.
[26] Xue Tianfan, Chen Baian, Wu Jiajun, et al. Video enhancement with task-oriented flow [J]. International Journal of Computer Vision, 2019, 127 (8): 1106-1125.
[27] Jo Y, Oh S W, Kang J, et al. Deep video super-resolution network using dynamic upsampling filters without explicit motion compensation [C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018: 3224-3232.
[28] Liu Ce, Sun Deqing. On Bayesian adaptive video super resolution [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2013, 36 (2): 346-360.
[29] Haris M, Shakhnarovich G, Ukita N. Recurrent back-projection network for video super-resolution [C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019: 3897-
[30] Yi Peng, Wang Zhongyuan, Jiang Kui, et al. Progressive fusion video super-resolution network via exploiting non-local spatio-temporal correlations [C]// Proceedings of IEEE/CVF International Conference on Computer Vision. 2019: 3106-3115.
[31] He Kaiming, Zhang Xiangyu, Ren Shaoqing, et al. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification [C]// Proceedings of the IEEE International Conference on Computer Vision. 2015: 1026-1034.
[32] Kingma D P, Ba J. Adam: A method for stochastic optimization [J]. arXiv preprint arXiv: 1412.6980, 2014.

[33] Li Wenbo, Tao Xin, Guo Taian, et al. Mucan: Multi-correspondence aggregation network for video super-resolution [C]// European Conference on Computer Vision. Springer, Cham, 2020: 335-351.
[34] Liu Ding, Wang Zhaowen, Fan Yuchen, et al. Learning temporal dynamics for video super-resolution: A deep learning approach [J]. IEEE Transactions on Image Processing, 2018, 27 (7): 3432-3445.
[35] Yi Peng, Wang Zhongyuan, Jiang Kui, et al. Progressive fusion video super-resolution network via exploiting non-local spatio-temporal correlations [C]// Proceedings of IEEE/CVF International Conference on Computer Vision. 2019: 3106-3115.
[36] Cao Jiezhang, Li Yawei, Zhang Kai, et al. Video super-resolution transformer [J]. arXiv preprint arXiv: 2106.06847, 2021.
[37] Dosovitskiy A, Beyer L, Kolesnikov A, et al. An image is worth 16x16 words: Transformers for image recognition at scale [J]. arXiv preprint arXiv: 2010.11929, 2020.

Note: Figure translations are in progress. See original paper for figures.