# Video Super-resolution with Temporal Group Attention

Takashi Isobe[1,2†], Songjiang Li[2], Xu Jia[2*], Shanxin Yuan[2], Gregory Slabaugh[2],
Chunjing Xu[2], Ya-Li Li[1], Shengjin Wang[1,*], Qi Tian[2]
[1]Department of Electronic Engineering, Tsinghua University
[2]Noah's Ark Lab, Huawei Technologies
jbj18@mails.tsinghua.edu.cn {liyali13, wgsgj}@tsinghua.edu.cn
{x.jia, songjiang.li, shanxin.yuan, gregory.slabaugh, tian.qi1}@huawei.com

## Abstract

*Video super-resolution, which aims at producing a high-resolution video from its corresponding low-resolution version, has recently drawn increasing attention. In this work, we propose a novel method that can effectively incorporate temporal information in a hierarchical way. The input sequence is divided into several groups, with each one corresponding to a kind of frame rate. These groups provide complementary information to recover missing details in the reference frame, which is further integrated with an attention module and a deep intra-group fusion module. In addition, a fast spatial alignment is proposed to handle videos with large motion. Extensive results demonstrate the capability of the proposed model in handling videos with various motion. It achieves favorable performance against state-of-the-art methods on several benchmark datasets. Code is available at https://github.com/junpan19/VSR_TGA.*

## 1. Introduction

Super-resolution aims at producing high-resolution (HR) images from the corresponding low-resolution (LR) ones by filling in missing details. For single image super-resolution, an HR image is estimated by exploring natural image priors and self-similarity within the image. For video super-resolution, both spatial information across positions and temporal information across frames can be used to enhance details for an LR frame. Recently the task of video super-resolution has drawn much attention in both the research and industrial communities. For example, video super-resolution is required when videos recorded for surveillance are zoomed in to recognize a person's identity or a car's license, or when videos are projected to a high definition display device for visually pleasant watching.


Figure 1. VSR results for the *Calender* clip in Vid4 [1]. Our method produces result with more details (cyan arrow), and fewer artifacts (red arrow) than DUF [8] and the recent proprosed EDVR [28].

Most video super-resolution methods [9, 1, 25, 29, 17] adopt the following pipeline: motion estimation, motion compensation, fusion and upsampling. They estimate optical flow between a reference frame and other frames in either an offline or online manner, and then align all other frames to the reference with backward warping. However, this is not optimal for video SR. Methods with explicit motion compensation rely heavily on the accuracy of motion estimation. Inaccurate motion estimation and alignment, especially when there is occlusion or complex motion, results in distortion and errors, deteriorating the final super-resolution performance. Besides, per-pixel motion estimation such as optical flow often suffers a heavy computational load. Recently Jo *et al.* [8] proposed the DUF method which implicitly utilizes motion information among LR frames to recover HR frames by means of dynamic upsampling filters. It is less influenced by the accuracy of motion estimation but its performance is limited by the size of the dynamic upsampling filters. In addition, the temporal information integration process from other frames to the reference frame is conducted without explicitly taking the reference frame into consideration. This

[†]The work was done in Noah's Ark Lab, Huawei Technologies.
[*]Corresponding author

leads to ineffective information integration for border frames in an input sequence.

In this work, we propose a novel deep neural network which hierarchically utilizes motion information in an implicit manner and is able to make full use of complementary information across frames to recover missing details for the reference frame. Instead of aligning all other frames to the reference frame with optical flow or applying 3D convolution to the whole sequence,we propose to divide a sequence into several groups and conduct information integration in a hierarchical way, that is, first integrating information in each group and then integrate information across groups. The proposed grouping method produces groups of subsequences with different frame rates, which provide different kinds of complementary information for the reference frame. Such different complementary information is modeled with an attention module and the groups are deeply fused with a 3D dense block and a 2D dense block to generate a high-resolution version of the reference frame. Overall, the proposed method follows a hierarchical manner. It is able to handle various kinds of motion and adaptively borrow information from groups of different frame rates. For example, if an object is occluded in one frame, the model would pay more attention to frames in which the object is not occluded.

However, the capability of the proposed method is still limited in dealing with video sequences of large motion since the receptive field is finite. To address this issue, a fast homography based method is proposed for rough motion compensation among frames. The resulting warped frames are not perfectly aligned but they suffer less distortion artifacts compared to existing optical flow based methods. Appearance difference among frames is indeed reduced such that the proposed neural network model can focus on object motion and produce better super-resolution result.

The proposed method is evaluated on several video super-resolution benchmarks and achieves state-of-the-art performance. We conduct further analysis to demonstrate its effectiveness.

To sum up, we make the following contributions:

- We propose a novel neural network which efficiently fuses spatio-temporal information through frame-rate-aware groups in a hierarchical manner.
- We introduce a fast spatial alignment method to handle videos with large motion.
- The proposed method achieves state-of-the-art performance on two popular VSR benchmarks.

## 2. Related Work

### 2.1. Single Image Super Resolution

Single image super-resolution (SISR) has benefited greatly from progress in deep learning. Dong [2] first proposed to use a three-layer CNN for SISR and showed impressive potential in super-resolving LR images. New architectures have been designed since then, including a very deep CNN with residual connections [10], a recursive architecture with skip-connections [11], a architecture with a sub-pixel layer and multi-channel output to directly work on LR images as input [23]. More recent networks, including EDSR [15], RDN [36], DBPN [4], RCAN [35], outperformed previous works by a large margin when trained on the novel large dataset DIV2K [27]. More discussions can be found in the recent survey [31].

### 2.2. Video Super Resolution

Video super resolution relies heavily on temporal alignment, either explicitly or implicitly, to make use of complementary information from neighboring low-resolution frames. VESCPN [1] is the first end-to-end video SR method that jointly trains optical flow estimation and spatial-temporal networks. SPMC [25] proposed a new sub-pixel motion compensation layer for inter-frame motion alignment, and achieved motion compensation and upsampling simultaneously. [29] proposed to jointly train the motion analysis and video super resolution in an end-to-end manner through a proposed task-oriented flow. [5] proposed to use a recurrent encoder-decoder module to exploit spatial and temporal information, where explicit inter-frame motion were estimated. Methods using implicit temporal alignment showed superior performance on several benchmarks. [12] exploited the 3DCNN's spatial-temporal feature representation capability to avoid motion alignment, and stacked several 3D convolutional layers for video SR. [8] proposed to use 3D convolutional layers to compute dynamic filters [7] for implicit motion compensation and upsampling. Instead of image level motion alignment, TDAN [26] and EDVR [28] worked in the feature level motion alignment. TDAN [26] proposed a temporal deformable alignment module to align features of different frames for better performance. EDVR [28] extended TDAN in two aspects by 1) using deformable alignment in a coarse-to-fine manner and 2) proposing a new temporal and spatial attention attention fusion module, instead of naively concatenating the aligned LR frames as TDAN does.

The work most related with ours is [17], which also reorganized the input frames to several groups. However, in [17], groups are composed of different number of input frames. In addition, that method generates an super-resolution result for each group and computes an attention map to combine these super-resolution results, which takes much computation and is not very effective. Our method divides input frames into several groups based on frame rate and effectively integrates temporal information in a hierarchical way.

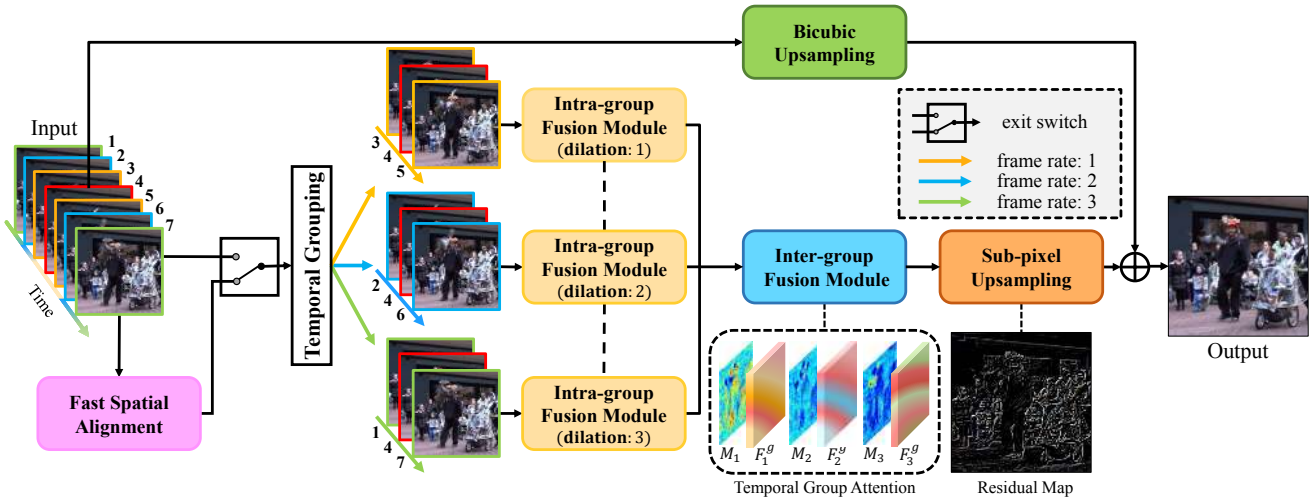



Figure 2. The proposed method with temporal group attention.

# 3. Methodology

## 3.1. Overview

Given a consecutive low-resolution video frame sequence consisting of one reference frame $I_t^L$ and $2N$ neighboring frames $\{I_{t-N}^L : I_{t-1}^L, I_{t+1}^L : I_{t+N}^L\}$, the goal of VSR is to reconstruct a high-resolution version of reference frame $\hat{I}_t$ by fully utilizing the spatio-temporal information across the sequence. The overall pipeline of the proposed method is shown in Fig. 2. It's a generic framework suitable for processing sequences of different input lengths. Take seven frames $\{I_1^L, I_2^L, ..., I_7^L\}$ for example, we denote the middle frame $I_4^L$ as the reference frame, and the other frames as neighboring ones. The seven input frames are divided into three groups based on decoupled motion, with each one representing a certain kind of frame rate. An intra-group fusion module with shared weights is proposed to extract and fuse spatio-temporal information within each group. Information across groups is further integrated through an attention-based inter-group fusion module. Finally, the output high-resolution frame $\hat{I}_4$ is generated by adding the network produced residual map and the bicubic upsampling of the input reference frame. Additionally, a fast spatial alignmemt module is proposed to further help deal with video sequences of large motion.

## 3.2. Temporal Group Attention

The crucial problem with implicit motion compensation lies on the inefficient fusion of temporal fusion in neighboring frames. In [8], input frames are stacked along the temporal axis and 3D convolutions are directly applied to the stacked frames. Such distant neighboring frames are not explicitly guided by the reference frame, resulting in insufficient information fusion, and this impedes the reference frame from borrowing information from distant frames. To address this issue, we propose to split neighboring $2N$ frames into N groups based on their temporal distances from the reference frame. Later, spatial-temporal information is extracted and fused in a hierarchical manner: an intra-group fusion module integrates information within each group, followed by an inter-group fusion module which effectively handles group-wise features.

**Temporal Grouping.** In contrast to the previous work, the neighboring $2N$ frames are split to $N$ groups based on the temporal distance to the reference frame. The original sequence is reordered as $\{G_1, ..., G_n\}$, $n \in [1 : N]$, where $G_n = \{I_{t-n}^L, I_t^L, I_{t+n}^L\}$ is a subsequence consisting of a former frame $I_{t-n}^L$, the reference frame $I_t^L$ and a latter frame $I_{t+n}^L$. Notice that the reference frame appears in each group. It is noteworthy that our method can be easily generalized to arbitrary frames as input. The grouping allows explicit and efficient integration of neighboring frames with different temporal distance for two reasons: 1) The contributions of neighboring frames in different temporal distances are not equal, especially for frames with large deformation, occlusion and motion blur. When a region in one group is (for example by occlusion), the missing information can be recovered by other groups. That is, information of different groups complements each other. 2) The reference frame in each group guides the model to extract beneficial information from neighboring frames, allowing efficient information extraction and fusion.

**Intra-group Fusion.** For each group, an intra-group fusion module is deployed for feature extraction and fusion within each group. The module consists of three parts. The first part contains three units as the spatial features extractor, where each unit is composed of a $3 \times 3$ convolutional layer followed by a batch normalization (BN) [6] and a ReLU [3]. All convolutional layers are equipped with dilation rate to model the motion level associated with a group. The dilation

rate is determined according to the frame rate in each group with the assumption that distant group has large motion and near group has small motion. Subsequently, for the second part, an additional 3D convolutional layer with $3 \times 3 \times 3$ kernel is used to perform spatio-temporal feature fusion. Finally, group-wise features $F_n^g$ are produced by applying eighteen 2D units in the 2D dense block to deeply integrate information within each group.

The weights of the intra-group fusion module are shared for each group for efficiency. The effectiveness of the proposed temporal grouping are presented in Sec.4.3.

**Inter-group Fusion with Temporal Attention.** To better integrate features from different groups, a temporal attention module is introduced. Temporal attention has been widely used in video-related tasks [24, 33, 34, 30]. In this work, we show that temporal attention also benefits the task of VSR by enabling the model to pay different attention across time. In the previous section, a frame sequence is categorized into groups according to different frame rates. These groups contain complementary information. Usually, a group with slow frame rate is more informative because the neighboring frames are more similar to the reference one. Simultaneously, groups with fast frame rate may also capture information about some fine details which are missing in the nearby frames. Hence, temporal attention works as a guidance to efficiently integrate features from different temporal interval groups.

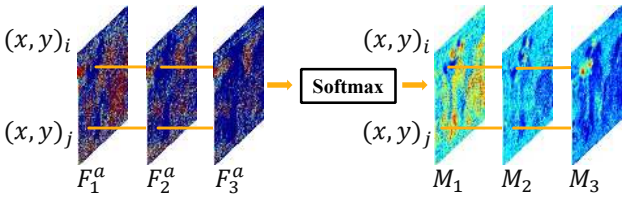



Figure 3. Computation of group attention maps. $F_n^a$ corresponds to group-wise features while $M_n$ is the attention mask.

For each group, a one-channel feature map $F_n^a$ is computed after applying a $3 \times 3$ convolutional layer on top of the corresponding feature maps $F_n^g$. They are further concatenated and a softmax function along temporal axis is applied to each position across channels to compute attention maps, as shown in Fig. 3. Each group's intermediate map is concatenated and the attention maps $M(x, y)$ are computed by applying softmax along temporal axis, as shown in Fig. 3.

$$M_n(x,y)_j = \frac{e^{F_n^a(x,y)_j}}{\sum_{i=1}^{N} e^{F_i^a(x,y)_j}} \tag{1}$$

Attention weighted feature for each group $\widetilde{F}_n^g$ is calculated as:

$$\widetilde{F}_n^g = M_n \odot F_n^g, n \in [1:N] \tag{2}$$

where $M_n(x,y)_j$ represents the weight of the temporal group attention mask at location $(x,y)_j$. $F_n^g$ represents the group-wise features produced by intra-group fusion module. '$\odot$' denotes element-wise multiplication.

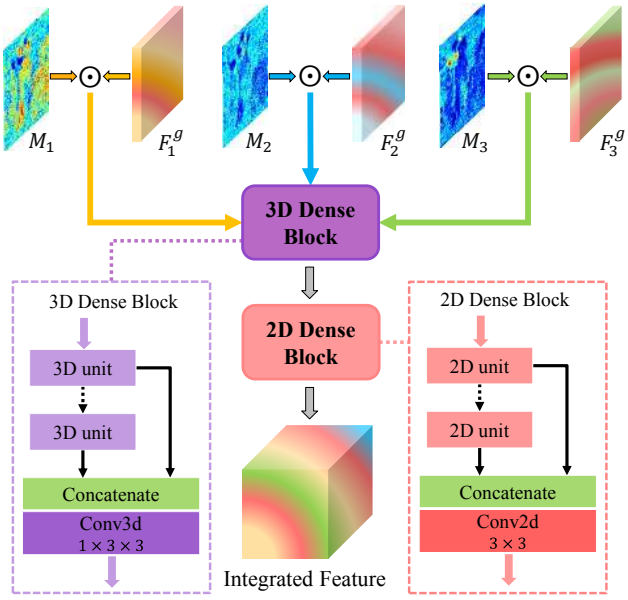



Figure 4. Structure of the inter-group fusion module.

The goal of the inter-group fusion module is to aggregate information across different temporal groups and produce a high-resolution residual map. In order to make full use of attention weighted feature over temporal groups, we first aggregate those features by concatenating them along the temporal axis and feed it into a 3D dense block. Then a 2D dense block is on top for further fusion, as shown in Fig. 4. 3D unit has the same structure as 2D unit which is used in intra-group fusion module. A convolution layer with $1 \times 3 \times 3$ kernel is inserted in the end of the 3D dense block to reduce channels. The design of 2D and 3D dense blocks are inspired by RDN [36] and DUF [8], which is modified in an efficient way to our pipeline.

Finally, similar to several single image super-resolution methods, sufficiently aggregated features are upsampled with a depth-to-space operation [23] to produce high-resolution residual map $R_t$. The high-resolution reconstruction $\hat{I}_t$ is computed as the sum of the residual map $R_t$ and a bicubic upsampled reference image $I_t^{\uparrow}$.

### 3.3. Fast Spatial Alignment

Although the proposed model is able to effectively use temporal information across frames, it has difficulty in dealing with videos with large motion. To improve the performance of the proposed model in case of large motion, we further propose a fast spatial alignment module. Different from previous methods [19, 1, 29] which either use offline optical flow or an integrated optical flow network for motion estimation and compensation, we estimate homography between every two consecutive frames and warp neighboring

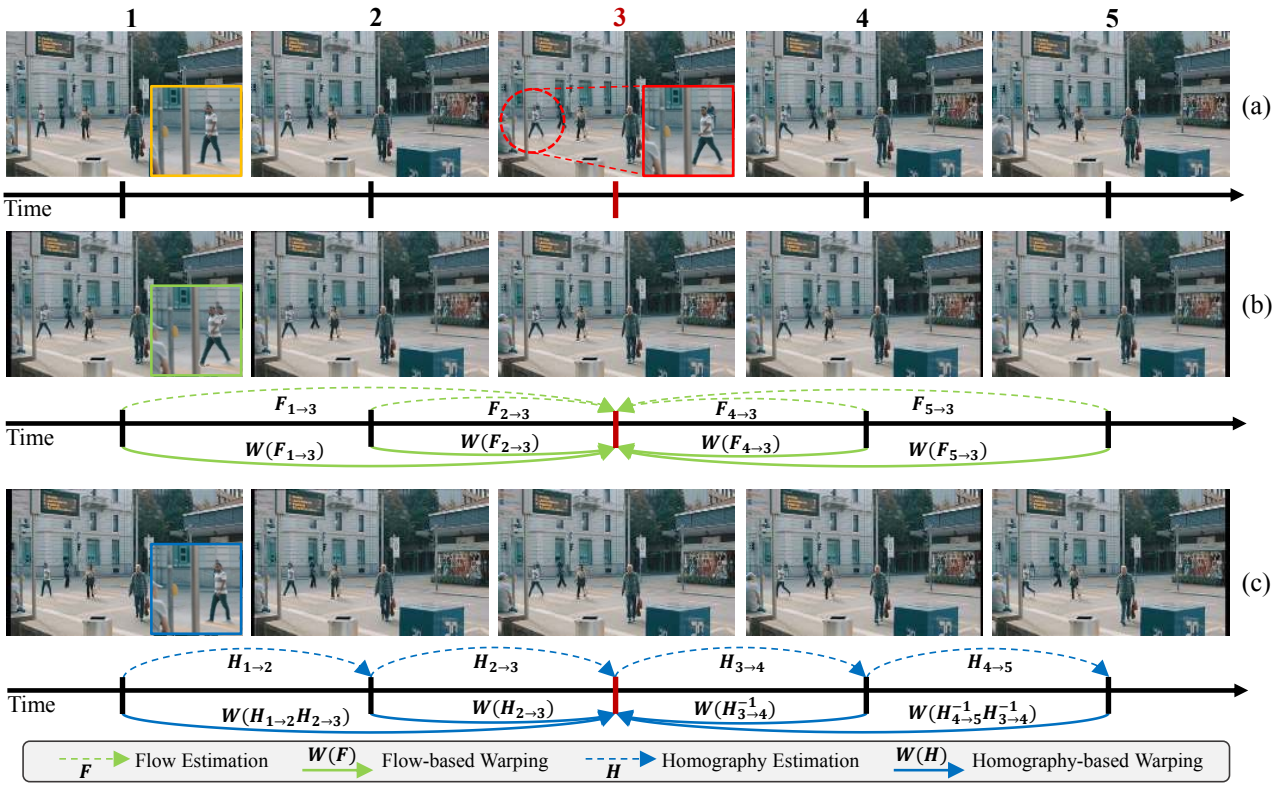



Figure 5. Fast spatial alignment compared with optical flow. (a) Original 5 consecutive frames, of which **frame 3** is the reference frame. (b) Alignment with optical flow. The flow for each neighboring frame is estimated independently. (c) The proposed alignment only estimates basic homographies for consecutive frames. The frame-level alignment suppresses pixel-level distortion. Zoom in for better visualization.

frames to the reference frame, which can be shown in Fig. 5. Interest points could be detected by feature detectors such as SIFT [18] or ORB [21], and point correspondences are computed to estimate homography. Homography from frame $A$ and $C$ can be computed as a product of the homography from $A$ to $B$ and the one from $B$ to $C$:

$$H_{A\to C} = H_{A\to B} \cdot H_{B\to C} \tag{3}$$

For a homography, the inverse transform can be represented by the inverse of the matrix:

$$H_{B\to A} = H_{A\to B}^{-1} \tag{4}$$

Since optical flow is computed for each pixel, imperfect optical flow estimation would introduce much unexpected pixel-level distortion into warping, destroying structure in original images. In addition, most optical-flow-based methods [14, 1, 25, 29] estimate optical flow between each neighboring frame and the reference frame independently, which would bring a lot of redundant computation when super-resolving a long sequence. In our method, since homography transformation is a global, it keeps the structure better and introduces little artifact. In addition, the associative composition nature of homography allows to decompose a homography between two frames into a product of homographies between every two consecutive ones in that interval, which avoids redundant computation and speeds up pre-alignment. Note that the pre-alignment here does not need to be perfect. As long as it does not introduce much pixel-level distortion, the proposed VSR network can give good performance. We also introduce exit mechanism for pre-alignment for robustness. That is, in case that few interest points are detected or there is much difference between a frame and the result after applying $H$ and $H^{-1}$, the frames are kept as they are without any pre-alignment. In other words, a conservative strategy is adopt in pre-alignment procedure.

## 4. Experiments

To evaluate the proposed method, a series of experiments are conducted and results are compared with existing state-of-the-art methods. Subsequently, a detailed ablation study is conducted to analyze the effectiveness of the proposed temporal grouping, group attention and fast spatial alignment. Results demonstrate the effectiveness and superiority of the proposed method.

| Method | # Frames | Calendar (Y) | City (Y) | Foliage (Y) | Walk (Y) | Average (Y) | Average (RGB) |
|---|---|---|---|---|---|---|---|
| Bicubic | 1 | 18.83/0.4936 | 23.84/0.5234 | 21.52/0.4438 | 23.01/0.7096 | 21.80/0.5426 | 20.37/0.5106 |
| SPMC † [25] | 3 | - | - | - | - | 25.52/0.76 | - |
| Liu† [17] | 5 | 21.61/ - | 26.29/ - | 24.99/ - | 28.06/ - | 25.23/ - | - |
| TOFlow [29] | 7 | 22.29/0.7273 | 26.79/0.7446 | 25.31/0.7118 | 29.02/0.8799 | 25.84/0.7659 | 24.39/0.7438 |
| FRVSR †[22] | recurrent | - | - | - | - | 26.69/0.822 | - |
| DUF-52L [8] | 7 | 24.17/0.8161 | 28.05/0.8235 | 26.42/0.7758 | 30.91/ **0.9165** | 27.38/0.8329 | **25.91/0.8166** |
| RBPN [5] | 7 | 24.02/0.8088 | 27.83/0.8045 | 26.21/0.7579 | 30.62/0.9111 | 27.17/0.8205 | 25.65/0.7997 |
| EDVR-L† [28] | 7 | 24.05/0.8147 | 28.00/0.8122 | 26.34/0.7635 | **31.02**/0.9152 | 27.35/0.8264 | 25.83/0.8077 |
| PFNL† [32] | 7 | **24.37/0.8246** | **28.09/0.8385** | **26.51/0.7768** | 30.65/0.9135 | **27.40/0.8384** | - |
| TGA (Ours) | 7 | **24.47/0.8286** | **28.37/0.8419** | **26.59/0.7793** | **30.96/0.9181** | **27.59/0.8419** | **26.10/0.8254** |

Table 1. Quantitative comparison (PSNR(dB) and SSIM) on **Vid4** for $4\times$ video super-resolution. Red text indicates the best and blue text indicates the second best performance. Y and RGB indicate the luminance and RGB channels, respectively. '†' means the values are taken from original publications or calculated by provided models. Best view in color.

| | Bicubic | TOFlow [29] | DUF-52L [8] | RBPN [5] | EDVR-L† [28] | TGA(Ours) |
|---|---|---|---|---|---|---|
| # Param. | N/A | 1.4M | 5.8M | 12.1M | 20.6M | 5.8M |
| FLOPs | N/A | 0.27T | 0.20T | 3.08T | 0.30T | 0.07T |
| Y Channel | 31.30/0.8687 | 34.62/0.9212 | 36.87/0.9447 | 37.20/0.9458 | **37.61/0.9489** | **37.59/0.9516** |
| RGB Channels | 29.77/0.8490 | 32.78/0.9040 | 34.96/0.9313 | 35.39/0.9340 | **35.79/0.9374** | **35.57/0.9387** |

Table 2. Quantitative comparison (PSNR(dB) and SSIM) on **Vimeo-90K-T** for $4\times$ video super-resolution. Red text indicaktes the best result and blue text indicates the second best. FLOPs are calculated on an LR image of size $112\times64$. '†' means the values are taken from original publications. Note that the deformation convolution and offline pre-alignment are not included in calculating FLOPs. Best view in color.

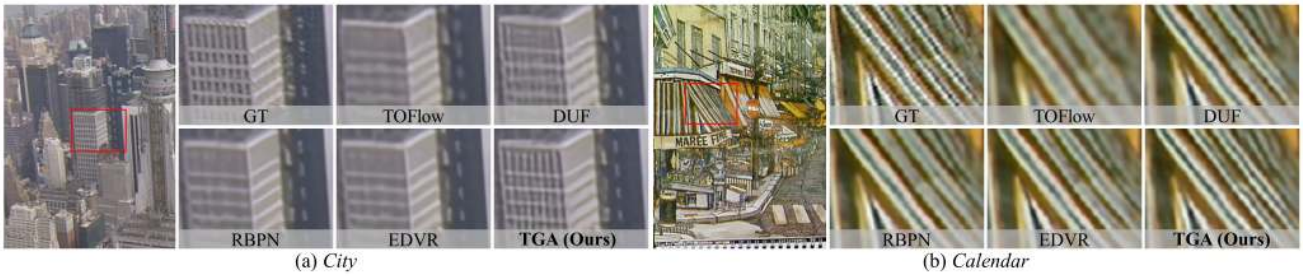
### 4.1. Implementation Details

**Dataset.** Similar to [5, 29], we adopt Vimeo-90k [29] as our training set, which is a widely used for the task of video super-resolution. We sample regions with spatial resolution $256\times256$ from high resolution video clips. Similar to [8, 29, 32] low-resolution patches of $64 \times 64$ are generated by applying a Gaussian blur with a standard deviation of $\sigma = 1.6$ and $4\times$ downsampling. We evaluate the proposed method on two popular benchmarks: Vid4 [16] and Vimeo-90K-T[29]. Vid4 consists of four scenes with various motion and occlusion. Vimeo-90K-T contains about $7k$ high-quality frames and diverse motion types.

**Implementation details.** In the intra-group fusion module, three 2D units are used for spatial features extractor, which is followed by a 3D convolution and eighteen 2D units in the 2D dense block to integrate information within each group. For the inter-group fusion module, we use four 3D units in the 3D dense block and twenty-one 2D units in the 2D dense block. The channel size is set to 16 for convolutional layers in the 2D and 3D units. Unless specified otherwise, our network takes seven low resolution frames as input. The model is supervised by pixel-wise $L1$ loss and optimized with Adam [13] optimizer in which $\beta_1 = 0.9$ and $\beta_2 = 0.999$. Weight decay is set to $5 \times 10^{-4}$ during training. The learning rate is initially set to $2 \times 10^{-3}$ and later down-scaled by a factor of 0.1 every 10 epoches until 30 epochs. The size of mini-batch is set to 64. The training data is augmented by flipping and rotating with a probability of 0.5. All experiments are conducted on a server with Python 3.6.4, PyTorch 1.1 and Nvidia Tesla V100 GPUs.

### 4.2. Comparison with State-of-the-arts

We compare the proposed method with six state-of-the-art VSR approaches, including TOFlow [29], SPMC [25], Liu [17], DUF [8], RBPN [5], EDVR [28] and PFNL [32]. Both TOFlow and SPMC apply explicit pixel-level motion compensation with optical flow estimation, while RBPN uses pre-computed optical flow as additional input. DUF, EDVR and PFNL conduct VSR with implicit motion compensation. We carefully implement TOFlow and DUF on our own, and rebuild RBPN and EDVR based on the publicly available code. We reproduce the performance of most of these methods as reported in the paper except for EDVR. Tab. 1 and Tab. 2 give quantitative results of state-of-the-art methods on Vid4 and Vimeo-90K-T, which are either reported in the original papers or computed by us. In the evaluation, we take all frames into account except for the DUF method [8] which crop 8 pixels on four borders of each frame since it suffer from severe border artifacts. In addition, we also include the number of parameters and FLOPs for most methods on an LR image of size $112 \times 64$ in Tab. 2. On Vid4 test set, the proposed method achieves a result of 27.59dB PSNR in the Y channel and 26.10dB PSNR in RGB channel, which outperforms other state-of-the-art methods by a large margin. Qualitative result in Fig. 6 also validates the superiority of the proposed method. Attributed to the proposed temporal group attention, which is able to make




(a) *City* (b) *Calendar*

Figure 6. Qualitative comparison on the **Vid4** for 4×SR. Zoom in for better visualization.


Figure 7. Qualitative comparison on the **Vimeo-90K-T** for 4×SR. Zoom in for better visualization.

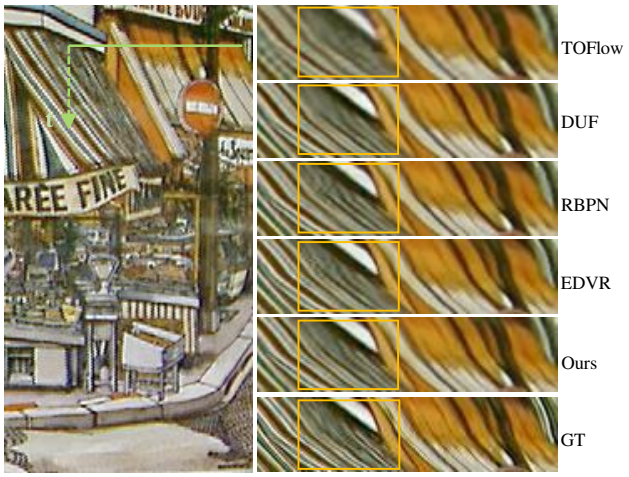



Figure 8. Visualization of temporal consistency for *calendar* sequence. Temporal profile is produced by recording a single pixel line (green line) spanning time and stacked vertically.

full use of complementary information among frames, our model produces sharper edges and finer detailed texture than other methods. In addition, we extract temporal profiles in order to evaluate the performance on temporal consistency in Fig.8. A temporal profile is produced by taking the same horizontal row of pixels from consecutive frames and stacking them vertically. The temporal profiles show that the proposed method gives temporally consistent results, which suffer less flickering artifacts than other approaches.

Vimeo-90K-T is a large and challenging dataset covering scenes with large motion and complicated illumination changes. The proposed method is compared with several methods including TOFlow, DUF, RBPN and EDVR. As shown in Tab. 2 and Fig. 7, the proposed method also achieves very good performance on this challenging dataset. It outperforms most state-of-the-art methods such as TOFlow, DUF and RBPN by a large margin both in PSNR and SSIM. The only exception is EDVR-L whose model size and computation is about four times larger than our method. In spite of this, our method is still rather comparable in PSNR and a little better in SSIM.

### 4.3. Ablation Study

In this section, we conduct several ablation study on the proposed temporal group attention and fast spatial alignment to further demonstrate the effectiveness of our method.

**Temporal Group Attention.** First we experiment with different ways of organizing the input sequence. One baseline method is to simply stack input frames along temporal axis and directly feed that to several 3D convolutional layers, similar to DUF [8]. Apart from our grouping method $\{345, 246, 147\}$, we also experiment with other ways of

| Model | DUF-like | $\{123, 345, 567\}$ | $\{345, 142, 647\}$ | $\{345, 246, 147\}$ |
|---|---|---|---|---|
| TG? | ✗ | ✓ | ✓ | ✓ |
| Vid4 | 27.18/0.8258 | 27.47/0.8384 | 27.54/0.8409 | **27.59/0.8419** |
| Vimeo-90K-T | 37.06/0.9465 | 37.46/0.9487 | 37.51/0.9509 | **37.59/0.9516** |

Table 3. Ablation on: different grouping strategies.

grouping: $\{123, 345, 567\}$ and $\{345, 142, 647\}$. As shown in Tab. 3, DUF-like input performs worst among these methods. That illustrate that integrating temporal information in a hierarchical manner is a more effective way in integrating information across frames. Both $\{345, 246, 147\}$ and $\{345, 142, 647\}$ are better than $\{123, 345, 567\}$, which implies the advantage of adding the reference frame in each group. Having the reference in the group encourages the model to extract complementary information that is missing in the reference frame. Another 0.05dB improvement of our grouping method $\{345, 246, 147\}$ could be attributed to the effectiveness of motion-based grouping in employing temporal information.

In addition, we also evaluate a model which removes the attention module from our whole model. As shown in Tab. 4, this model performs a little worse than our full model. We also train our full model with a sequence of 5 frames as input. The result in Tab. 4 shows that the proposed method can effectively borrow information from additional frames. We notice that the proposed method outperforms DUF even with 2 fewer frames in the input. In addition, we conduct a toy experiment where a part of a neighboring frame is occluded and visualize the maps of temporal group attention. As shown in Fig. 9, the model does attempt to borrow more information from other groups when a group can not provide complementary information to recover the details of that region.

| Model | Model 1 | Model 2 | Model 3 |
|---|---|---|---|
| # Frames | 7 | 5 | 7 |
| GA? | ✗ | ✓ | ✓ |
| Vid4 | 27.51/0.8394 | 27.39/0.8337 | **27.59/0.8419** |
| Vimeo-90K-T | 37.43/0.9506 | 37.34/0.9491 | **37.59/0.9516** |

Table 4. Ablations on: group attention (GA) modules and the influence of the different input frames in our hierarchical information aggregation way.

**Fast Spatial Alignment.** To investigate the effectiveness and efficiency of the proposed fast spatial alignment, we equip the proposed TGA model with three different pre-alignment strategies: TGA without alignment, TGA with PyFlow [20], and TGA with FSA. The evaluation is conducted on Vimeo-90K-T where there is various motion in the video clips. Tab. 5 shows the performance of TGA with pyflow is significantly inferior than the TGA model without any pre-alignment. It implies that imperfect optical flow estimation leads to inaccurate motion compensation such as distortion on the regions with large motion (see the green box in Fig. 5), which confuses the model during training and hurts the final video super-resolution performance. In contrast, the proposed FSA boosts the performance of the TGA model from 37.32dB to 37.59dB. This demonstrates that the proposed FSA, which although does not perfectly align frames, is capable of reducing appearance differences among frames in a proper way. We also compute time cost of this module on Vimeo-90K-T dataset and present it in Tab. 5. Our FSA method is much more efficient than the PyFlow method. Note that since every sequence in Vimeo-90K-T only contains 7 frames, the advantage of FSA in reducing redundant computation is not fully exployed. Both PyFlow and our FSA are run on CPU, and FSA could be further accelerated with optimized GPU implementation.

| Pre-alignment | w/o | w/ PyFlow [20] | w/ FSA |
|---|---|---|---|
| PSNR/SSIM | 37.32/0.9482 | 35.14/0.9222 | **37.59/0.9516** |
| Time (CPU+GPU) | 0+70.8ms | 760.2+70.8ms | 18.6+70.8ms |

Table 5. Ablation on: the effectiveness and efficiency of the fast spatial alignment module. The elapsed time are calculated on processing a seven frame sequence with LR size of 112×64.


Figure 9. Visualization of group attention masks under occlusion settings. $G_1, G_2$ and $G_3$ denote three groups.

## 5. Conclusion

In this work, we proposed a novel deep neural network which hierarchically integrates temporal information in an implicit manner. To effectively leverage complementary information across frames, the input sequence is reorganized into several groups of subsequences with different frame rates. The grouping allows to extract spatio-temporal information in a hierarchical manner, which is followed by an intra-group fusion module and inter-group fusion module. The intra-group fusion module extracts features within each group, while the inter-group fusion module borrows complementary information adaptively from different groups. Furthermore, an fast spatial alignment is proposed to deal with videos in case of large motion. The proposed method is able to reconstruct high-quality HR frames and also maintain the temporal consistency. Extensive experiments on several benchmark datasets demonstrate the effectiveness of the proposed method.

## References


[1] Jose Caballero, Christian Ledig, Andrew Aitken, Alejandro Acosta, Johannes Totz, Zehan Wang, and Wenzhe Shi. Real-time video super-resolution with spatio-temporal networks and motion compensation. In *CVPR*, 2017.

[2] Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang. Learning a deep convolutional network for image super-resolution. In *ECCV*, 2014.

[3] Xavier Glorot, Antoine Bordes, and Yoshua Bengio. Deep sparse rectifier neural networks. In *AISTATS*, 2011.

[4] Muhammad Haris, Gregory Shakhnarovich, and Norimichi Ukita. Deep back-projection networks for super-resolution. In *CVPR*, 2018.

[5] Muhammad Haris, Gregory Shakhnarovich, and Norimichi Ukita. Recurrent back-projection network for video super-resolution. In *CVPR*, 2019.

[6] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *ICML*, 2015.

[7] Xu Jia, Bert De Brabandere, Tinne Tuytelaars, and Luc V Gool. Dynamic filter networks. In *NeurIPS*, 2016.

[8] Younghyun Jo, Seoung Wug Oh, Jaeyeon Kang, and Seon Joo Kim. Deep video super-resolution network using dynamic upsampling filters without explicit motion compensation. In *CVPR*, 2018.

[9] Armin Kappeler, Seunghwan Yoo, Qiqin Dai, and Aggelos K Katsaggelos. Video super-resolution with convolutional neural networks. *IEEE Transactions on Computational Imaging*, 2(2):109–122, 2016.

[10] Jiwon Kim, Jung Kwon Lee, and Kyoung Mu Lee. Accurate image super-resolution using very deep convolutional networks. In *CVPR*, 2016.

[11] Jiwon Kim, Jung Kwon Lee, and Kyoung Mu Lee. Deeply-recursive convolutional network for image super-resolution. In *CVPR*, 2016.

[12] Soo Ye Kim, Jeongyeon Lim, Taeyoung Na, and Munchurl Kim. 3dsrnet: Video super-resolution using 3d convolutional neural networks. *CoRR*, abs/1812.09079, 2018.

[13] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015.

[14] Renjie Liao, Xin Tao, Ruiyu Li, Ziyang Ma, and Jiaya Jia. Video super-resolution via deep draft-ensemble learning. In *ICCV*, 2015.

[15] Bee Lim, Sanghyun Son, Heewon Kim, Seungjun Nah, and Kyoung Mu Lee. Enhanced deep residual networks for single image super-resolution. In *CVPR Workshops*, 2017.

[16] Ce Liu and Deqing Sun. On bayesian adaptive video super resolution. *IEEE transactions on pattern analysis and machine intelligence*, 36(2):346–360, 2013.

[17] Ding Liu, Zhaowen Wang, Yuchen Fan, Xianming Liu, Zhangyang Wang, Shiyu Chang, and Thomas Huang. Robust video super-resolution with learned temporal dynamics. In *ICCV*, 2017.

[18] David G Lowe. Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, 60(2):91–110, 2004.

[19] Ziyang Ma, Renjie Liao, Xin Tao, Li Xu, Jiaya Jia, and Enhua Wu. Handling motion blur in multi-frame super-resolution. In *CVPR*, 2015.

[20] Deepak Pathak, Ross Girshick, Piotr Dollár, Trevor Darrell, and Bharath Hariharan. Learning features by watching objects move. In *CVPR*, 2017.

[21] Ethan Rublee, Vincent Rabaud, Kurt Konolige, and Gary R Bradski. Orb: An efficient alternative to sift or surf. In *ICCV*, 2011.

[22] Mehdi SM Sajjadi, Raviteja Vemulapalli, and Matthew Brown. Frame-recurrent video super-resolution. In *CVPR*, 2018.

[23] Wenzhe Shi, Jose Caballero, Ferenc Huszár, Johannes Totz, Andrew P Aitken, Rob Bishop, Daniel Rueckert, and Zehan Wang. Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network. In *CVPR*, 2016.

[24] Sijie Song, Cuiling Lan, Junliang Xing, Wenjun Zeng, and Jiaying Liu. An end-to-end spatio-temporal attention model for human action recognition from skeleton data. In *AISTATS*, 2017.

[25] Xin Tao, Hongyun Gao, Renjie Liao, Jue Wang, and Jiaya Jia. Detail-revealing deep video super-resolution. In *ICCV*, 2017.

[26] Yapeng Tian, Yulun Zhang, Yun Fu, and Chenliang Xu. Tdan: Temporally deformable alignment network for video super-resolution. *CoRR*, abs/1812.02898, 2018.

[27] Radu Timofte, Eirikur Agustsson, Luc Van Gool, Ming-Hsuan Yang, and Lei Zhang. Ntire 2017 challenge on single image super-resolution: Methods and results. In *CVPR Workshops*, 2017.

[28] Xintao Wang, Kelvin CK Chan, Ke Yu, Chao Dong, and Chen Change Loy. Edvr: Video restoration with enhanced deformable convolutional networks. In *CVPR Workshops*, 2019.

[29] Tianfan Xue, Baian Chen, Jiajun Wu, Donglai Wei, and William T Freeman. Video enhancement with task-oriented flow. *International Journal of Computer Vision*, 127(8):1106–1125, 2019.

[30] Chenggang Yan, Yunbin Tu, Xingzheng Wang, Yongbing Zhang, Xinhong Hao, Yongdong Zhang, and Qionghai Dai. Stat: spatial-temporal attention mechanism for video captioning. *IEEE Transactions on Multimedia*, 2019.

[31] Wenming Yang, Xuechen Zhang, Yapeng Tian, Wei Wang, Jing-Hao Xue, and Qingmin Liao. Deep learning for single image super-resolution: A brief review. *IEEE Transactions on Multimedia*, 2019.

[32] Peng Yi, Zhongyuan Wang, Kui Jiang, Junjun Jiang, and Jiayi Ma. Progressive fusion video super-resolution network via exploiting non-local spatio-temporal correlations. In *ICCV*, 2019.

[33] Mihai Zanfir, Elisabeta Marinoiu, and Cristian Sminchisescu. Spatio-temporal attention models for grounded video captioning. In *ACCV*, 2016.

[34] Jinliang Zang, Le Wang, Ziyi Liu, Qilin Zhang, Gang Hua, and Nanning Zheng. Attention-based temporal weighted convolutional neural network for action recognition. In *IFIP*, 2018.

[35] Yulun Zhang, Kunpeng Li, Kai Li, Lichen Wang, Bineng Zhong, and Yun Fu. Image super-resolution using very deep residual channel attention networks. In *ECCV*, 2018.

[36] Yulun Zhang, Yapeng Tian, Yu Kong, Bineng Zhong, and Yun Fu. Residual dense network for image super-resolution. In *CVPR*, 2018.