Local-Global Fusion Network for Video Super-Resolution

DEWEI SU†, HUA WANG‡, LONGCUN JIN†, (Member, IEEE), XIANFANG SUN‡, AND XINYI PENG†

†School of Software Engineering, South China University of Technology, Guangzhou 510006, China
‡School of Computer Science and Informatics, Cardiff University, Cardiff CF24 3AA, U.K.

Corresponding author: Longcun Jin (lcjin@scut.edu.cn)

This work was supported in part by the Pearl River Technology Nova Project under Grant 201710010020, in part by the Fundamental Research Funds for the Central Universities under Grant 2019MS086, and in part by the National Science Foundation of China under Grant 61300135.

ABSTRACT The goal of video super-resolution technique is to address the problem of effectively restoring high-resolution (HR) videos from low-resolution (LR) ones. Previous methods commonly used optical flow to perform frame alignment and designed a framework from the perspective of space and time. However, inaccurate optical flow estimation may occur easily which leads to inferior restoration effects. In addition, how to effectively fuse the features of various video frames remains a challenging problem. In this paper, we propose a Local-Global Fusion Network (LGFN) to solve the above issues from a novel viewpoint. As an alternative to optical flow, deformable convolutions (DCs) with decreased multi-dilation convolution units (DMDCUs) are applied for efficient implicit alignment. Moreover, a structure with two branches, consisting of a Local Fusion Module (LFM) and a Global Fusion Module (GFM), is proposed to combine information from two different aspects. Specifically, LFM focuses on the relationship between adjacent frames and maintains the temporal consistency while GFM attempts to take advantage of all related features globally with a video shuffle strategy. Benefiting from our advanced network, experimental results on several datasets demonstrate that our LGFN can not only achieve comparative performance with state-of-the-art methods but also possess reliable ability on restoring a variety of video frames. The results on benchmark datasets of our LGFN are presented on https://github.com/BIOINSu/LGFN and the source code will be released as soon as the paper is accepted.

INDEX TERMS Convolutional neural networks, deep learning, improved deformable convolution, local-global feature fusion, video super-resolution.

I. INTRODUCTION

As one of the fundamental sub-tasks of video enhancement, video super-resolution (VSR) aims at mapping the low-resolution (LR) videos into corresponding high-resolution (HR) ones. VSR can either be regarded as a separate mission or be combined with other works, such as video interpolation [1]. In addition, high availability of various videos makes it practical in many scenarios including satellite [2], high dynamic range (HDR) [3] and surveillance videos [4], [5]. Therefore, research on VSR has both scientific and realistic significance.

It is widely known that super-resolution (SR) problem is ill-posed and VSR is no exception. Such characteristic means that one given input may correspond to various results and how to restore a sophisticated output still remains difficult. On the other hand, while single image super-resolution (SISR) had a breakthrough in recent years, directly applying SISR on multi-frames with time dependency can result in neglect of temporal relationship. By taking advantage of correlated video frames, features can complement with each other to obtain better reconstruction outcomes and maintain temporal consistency at the same time. Similar studies which use multiple images for reconstruction also appears in stereo image SR [6]–[9], light field SR [10]–[12] and multi-image SR (MISR) [13]–[16]. The connections between images in these SR tasks are mostly related to different viewpoints or small spacial displacements instead of time.

Since existing approaches [17]–[31] commonly take several LR frames as input and reconstruct only the central
reference frame, one key issue need to be tackled is how to effectively use information from neighboring frames. A lot of works [24]–[30] have proved that aligning neighboring frames with the reference frame is a feasible choice rather than directly concatenate them together. Benefiting from the development of optical flow estimation methods, making use of optical flow for warping and motion compensation is a useful strategy [24]–[27]. However, it is hard to obtain accurate prediction using optical flow due to motion blur and it might also introduce unnecessary noise to the final SR results. Besides, such tactics usually require high computational cost.

Another inevitable question is how to reasonably fuse features together to achieve better results. One simple and straightforward way is to splice all the features together and leave the rest of the framework to learn fusion strategy automatically [17], [18]. However, stacking numerous information can cause great pressure on the following components, making the model unable to fully utilize effective features. In order to overcome such drawbacks, several methods [21]–[23] attempted to integrate the current features with one neighbor in front and one behind, while some others [29]–[31] devoted to combine arbitrary neighboring frames with the reference frame directly. The former imposes a strong coherence between adjacent features and the latter involves abundant temporal information to complement the central features. Within our fusion module, we cover the two aspects to ensemble both advantages, so as to acquire adequately fused features.

In this paper, we propose a Local-Global Fusion Network (LGFN) for restoring the degraded videos into a high-quality state. Overall, our method can be divided into the following stages. Firstly, a shallow feature extractor is applied to the input frames. Then, we align the neighboring frames with the reference frame at the feature level through stacked and improved deformable convolutions (DCs). In contrast to the normal operation, in each DC, we utilize a decreased multi-dilation convolution unit (DMDCU) to predict more accurate sampling offsets and modulation scalars. Specifically, the number of dilation convolutions and the dilation rates in each DMDCU taper off according to the position of the DC. The deeper, the less. This idea is based on the observation that shallow neighboring features, which have more displacements with the reference features, need larger receptive fields to capture the motions between pixels, whereas deep features that have already been roughly aligned are easier to achieve accurate alignment. Such practice not only improves performance but also maintains a small increase on computational cost. After the alignment, the features of consecutive frames are sent into two different branches respectively for feature fusion. One of the branches, called Local Fusion Module (LFM), is designed to combine information between adjacent frames so that the generated features can maintain close correlation and temporal consistency. In order to exploit abundant features from all input frames, the other branch, called Global Fusion Module (GFM), fuses the features by making use of a video shuffle [32] strategy across all frames. The information can be exchanged globally, which brings more available and effective features for each frame. Next, the features from both branches are integrated correspondingly by a 1 × 1 convolution and then merge together to represent the features of the central reference frame. In the end, SR reconstruction module is responsible for restoring high-resolution video frames, where we apply the advanced residual channel attention blocks (RCABs) to fulfill the task.

Overall, the major contributions of this work can be summarized as follows:

- We present an end-to-end trainable deep learning model LGFN for high-quality video super-resolution. Experimental results on several benchmark datasets demonstrate that our concise and effective network can achieve comparative performance with the state-of-the-art methods.
- In order to proceed motion compensation implicitly, we introduce an optimized deformable convolution (DC) with a decreased multi-dilation convolution unit (DMDCU). Unlike the ordinary version of DC, the modification can predict more accurate parameters and thus perform feature alignment more precisely.
- We carefully design a structure with two branches for superior feature fusion. Local Fusion Module LFM imposes strong correlation to adjacent features and maintains temporal consistency. Global Fusion Module GFM exchanges information among all features and utilizes more abundant details through a video shuffle strategy, which mutually reinforces the fusion capability. Both branches dedicate to performing feature fusion effectively.

The rest of this paper is structured as follows. Section 2 presents the related work including SISR and VSR. Section 3 explains the network architecture and introduces each component in detail. Experimental results with both quantitative and visual comparisons are shown in Section 4. Ablation experiments in Section 5 study the specific impact of each module. Finally, conclusion and future work are discussed in Section 6.

II. RELATED WORK

A. SINGLE IMAGE SUPER-RESOLUTION

Single image super-resolution (SISR) has achieved impressive performance in the past few years. Dong et al. [33] first utilized a simple network SRCNN for super-resolution, which only consists of three convolution layers. They upscaled the LR images at the front of the network, and some subsequent approaches inherited this strategy. With the network becomes deeper, such practice results in high computational cost for the following components. To overcome such limitations, FSRCNN [34] applied a deconvolution layer at the end of the network for upsampling so that the middle mapping layers can be shared for different resolutions. Another efficient design ESPCN [35] used a sub-pixel convolution layer which transforms channel dimension into space dimension. Thus the upsampling process can be learned implicitly.
More complex and advanced models are introduced in recent study. Tong et al. [36] put forward SRDenseNet with dense blocks which utilizes features from previous layers by concatenation. RDN [37] combines the advantages of both residual learning and skip connections to achieve better results. Dai et al. [38] came up with an attention skill called second-order attention which focuses on the feature correlation inside layers. Zhang et al. [39] proposed the residual channel attention network (RCAN) with attention in channel dimension to obtain amplitude high frequency information. We will utilize the basic structure of RCAN in our reconstruction module.

B. VIDEO SUPER-RESOLUTION

Since video is a kind of sequence, it is feasible to use a recurrent structure to model the temporal information. Sajjadi et al. [20] introduced the frame-recurrent video super-resolution (FRVSR) approach to restore an HR frame from the current and the former LR frames, where the former one is warped using current predicted optical flow. Yan et al. [21] proposed the frame and feature-context video super-resolution (FFCVSR) approach, where the former information is replaced with the current ones after several iterations to reduce the amassing deviation. Zhu et al. [22] put forward the residual invertible spatio-temporal network (RISTN) for video super-resolution, which consists of a residual dense connected LSTM to capture the dependencies among frames. Fuoli et al. [23] further utilized the recurrent latent space propagation (RLSP) with high-dimensional latent states to efficiently restore video frames. However, once it comes to using hidden states to save previous information, such strategy makes the restoration effect of the first few frames relatively poor, since there is no corresponding prior information.

The usage of optical flow estimation for explicit motion compensation is a prevalent approach for VSR. Tao et al. [24] proposed a sub-pixel motion compensation (SPMC) layer and performed alignment with optical flow. Wang et al. [25] came up with a multi-memory convolutional neural network (MMCNN), which jointly trains an optical flow estimation network and a multi-memory reconstruction network. Xue et al. [26] introduced TOFlow, which produces task-oriented optical flows through a gaussian pyramid architecture. The flows are gradually refined and warped later only once through a spatial transformer network. Li et al. [27] proposed a deep dual attention network (DDAN), which applies a pyramid multi-level structure to predict optical flows and performs warping on every resolution. Yi et al. [40] utilized optical flow for simple motion compensation and proposed an ultra dense memory block (UDMB), which combines hierarchical features to improve performance. Wang et al. [41] considered that there is a gap between the LR optical flows and the HR outputs, which could result in loss of details. Thus, they proposed super-resolve optical flows for video SR (SOF-VSR), which predicts the optical flows at high-resolution scales to reduce the displacement. Although optical flow is an effective means of motion compensation, it is challenging to obtain accurate optical flow estimation results. In addition, the computational cost and the introduction of extra noise are noteworthy problems, which may result in undesired artifacts.

To overcome the limitations of optical flow estimation, several methods provided valuable directions. Jo et al. [31] proposed to use dynamic upsampling filters (DUFs) on input center frames for reconstruction. Liu et al. [28] put forward a locally connected video super-resolution (LCVSR) approach, which also utilizes a similar tactic to perform motion compensation and later refines the features through a refinement network. Furthermore, Haris et al. [29] proposed a recurrent back-projection network (RBPN) for video super-resolution, which adopts an encoder-decoder based module to continuously fuse information between frames.

Recently, Tian et al. [30] proposed a temporally-deformable alignment network (TDAN), which takes advantage of deformable convolution [42] for implicit frame alignment and achieves splendid performance. Since the sampling point of the kernel can be shifted on the input feature maps, deformable convolution can automatically learn the effective features of the neighboring frames and offer valid information for reconstruction of the reference frame. However, using ordinary convolution to predict the required parameters may limit the relevant performance. Thus, in our work, we will utilize the deformable convolution v2 [43] for the alignment module, and further improve the performance by our proposed decreased multi-dilation convolution unit (DMDCU), which obtains more accurate results and keeps the extra computational cost relatively low.

III. LOCAL-GLOBAL FUSION NETWORKS

A. NETWORK ARCHITECTURE

In this section, we first introduce the overall architecture of our method and then describe each component in detail. The input of our network is a sequence of low-resolution video frames \( \{I_{LR}^{t-N}, I_{LR}^{t-1}, I_{LR}^{t}, I_{LR}^{t+1}, \ldots, I_{LR}^{t+N}\} \), where \( I_{LR}^{t} \) is the reference frame while the others are the neighboring frames. For VSR, our purpose is to reconstruct a high-resolution reference frame \( I_{SR}^{t} \) through fully exploiting the correlations among the video frames. The framework of our approach is depicted in Fig. 1 and the pseudo-code of the training phase is shown in Algorithm 1. The whole process is described as follows.

The input frames as RGB images are first fed into a feature extractor, which consists of a simple convolution layer and a number of residual blocks, to extract shallow features. The process can be written as:

\[
F_T = \text{Net}_{\text{feats}}(I_{LR}^{t})
\]  

where \( F_T \) denotes the extracted features and \( \text{Net}_{\text{feats}} \) is the feature extractor.

The extracted neighboring features \( F_t \) are then aligned to the reference feature \( F_t \) through the alignment module, which consists of stacked deformable convolutions with decreased
multi-dilation convolution units. The implicit motion compensation in the alignment module can be represented as:

\[ F_i^a = Net_{align}(F_i), \quad i \in [t-N, t+N] \quad \text{and} \quad i \neq t \quad (2) \]

where \( F_i^a \) is the aligned neighboring features.

Following the alignment, the features go through two branches to fuse two different types of information. One branch is the local fusion module (LFM) which focuses on the correlation between adjacent frames so as to maintain the temporal consistency. The LFM can be formulated as:

\[ \{L_{t-N}, \ldots, L_{t}, \ldots, L_{t+N}\} = Net_{local}(\{F_i^a, \ldots, F_{i-N}, \ldots, F_{i+N}\}) \quad (3) \]

where \( \{L_{t-N}, \ldots, L_{t}, \ldots, L_{t+N}\} \) represents the locally fused features and \( \{F_i^a, \ldots, F_{i-N}, \ldots, F_{i+N}\} \) are the aligned features. In the meantime, the other branch, called global fusion module (GFM), aiming at sufficiently utilizing the information of all the input frames, produces the global fused features \( \{G_{t-N}, \ldots, G_{t}, \ldots, G_{t+N}\} \), which are calculated by:

\[ \{G_{t-N}, \ldots, G_{t}, \ldots, G_{t+N}\} = Net_{global}(\{F_i^a, \ldots, F_{i-N}, \ldots, F_{i+N}\}) \quad (4) \]

Afterwards, the features from both branches are concatenated and merged together through a simple 1 \times 1 convolution:

\[ F_i^M = Conv_1(L_i, G_i), \quad i \in [t-N, t+N] \quad (5) \]

where \( F_i^M \) is the integrated output.

Next, the merged features pass through a 3 \times 3 convolution layer for feature fusion and channel reduction. Then, the advanced residual channel attention blocks (RCABs) are applied for high performance SR reconstruction. A skipped connection from shallow features of reference frame is used for residual learning. The whole process can be defined as:

\[ F_{fus} = Conv_3(F_i^M, \ldots, F_{i-N}^M, \ldots, F_{i+N}^M) \quad (6) \]

\[ F_i^Rec = Net_{RCABs}(F_{fus}) + F_i \quad (7) \]

In the end, the output is sent to a pixel shuffle layer for upsampling and then added into the bilinear upsampled LR frame from a skipped connection to construct the final super-resolved reference frame:

\[ I_i^{SR} = Net_{up}(F_i^{Rec}) + \text{bilinear}(I_t^{LR}) \quad (8) \]

where \( I_i^{SR} \) is the restored high-resolution reference frame.

### B. ALIGNMENT MODULE

We now introduce the alignment module in detail. For the purpose of fully utilizing the neighboring frames, methods based on optical flow warping can suffer from undesired artifacts and expensive computation. To overcome these shortcomings, we stack the superior modulated deformable convolution [43] to perform frame alignment and implicit motion compensation.

Common convolution operation can be described as:

\[ y(p) = \sum_{k=1}^{K} w_k \cdot x(p + p_k) \quad (9) \]

Here, \( y(p) \) refers to the output features \( y \) on each position \( p \), \( w_k \) is the weight, and \( p_k \) is the location offset on the kernel sampling grid. For instance, if the convolution kernel is 3 \times 3, \( K \) equals to 9, and \( p_k \in \{(-1, -1), (-1, 0), \ldots, (1, 0), (1, 1)\} \). Essentially, modulated deformable convolution adds predictable offsets and modulation scalars on the original fixed sampling position. As depicted in Fig.2, such tactic can adaptively learn dynamic sampling positions on the neighboring features that need to be aligned with the reference features. Let \( F_{i}^{b-1} \) denote the input neighboring features of the \( b \)-th modulated deformable convolution. The aligned output \( F_i^b \) can be calculated as:

\[ F_i^b(p) = \sum_{k=1}^{K} w_k \cdot F_i^{b-1}(p + p_k + \Delta p_{i,k}) \cdot \Delta m_{i,k} \quad (10) \]
Algorithm 1 Local-Global Fusion Network

Input: Training set \( \{ I_{HR}^L, I_{HR}^R \} \).
Input: Iteration epoch \( E \), mini-batch size \( b \), number of input frames \( 2N + 1 \) and learning rate \( \eta \).
Output: Super-resolved reference frame \( I_{SR}^L \).

1. Initialize network parameter \( \Theta \).
2. for each epoch \( e \) in \( E \) do
   
   /* Shallow Feature Extraction */
   
   \( F_T \leftarrow \text{Net}_{\text{feat}}(I_T^{LR}) \)

   /* Alignment */
   
   \( F_i \leftarrow \text{Net}_{\text{align}}(F_i, F_t) \), \( i \in [t - N, t + N] \) and \( i \neq t \)

   /* Feature Fusion */
   
   \( \{ L_{t-N}, \ldots, L_t, \ldots, L_{t+N} \} \leftarrow \text{Net}_{\text{local}}(F_{t-N}^a, \ldots, F_t^a, \ldots, F_{t+N}^a) \)

   \( \{ G_{t-N}, \ldots, G_t, \ldots, G_{t+N} \} \leftarrow \text{Net}_{\text{global}}(F_{t-N}^a, \ldots, F_t^a, \ldots, F_{t+N}^a) \)

   \( F_i^M \leftarrow \text{Conv}_{1}([L_i, G_i]), \) \( i \in [t - N, t + N] \)

   /* Reconstruction */
   
   \( F_{\text{fus}} \leftarrow \text{Conv}_{3}([F_{t-N}^M, \ldots, F_t^M, \ldots, F_{t+N}^M]) \)

   \( F_{\text{Rec}} \leftarrow \text{Net}_{\text{RCAB}}(F_{\text{fus}}) + F_t \)

   \( I_t^{SR} \leftarrow \text{Net}_{\text{up}}(F_{\text{Rec}}) + \text{bilinear}(I_t^{LR}) \)

   /* Loss Computation and Optimization */
   
   12: Training loss \( L_{\Theta} = \| I_{t}^{SR} - I_{t}^{HR} \|_p \)

   13: Update network parameter \( \Theta \) with ADAM optimizer: \( \Theta \leftarrow \text{Optimizer}(L_{\Theta}, \eta) \)

14: end for
15: end for

where \( \Delta p_i, k \) and \( \Delta m_i, k \) are the offset and the modulation scalar of the \( k \)-th sampling location. Bilinear interpolation is applied to compute \( F_i^{p-1}(p + p_k + \Delta p_i, k) \) in case that \( \Delta p_i, k \) is fractional. Note that both parameters are learned automatically and depend on the relationship between the neighboring features \( F_i \) and the reference features \( F_t \). The process of obtaining deformable parameters can be expressed as:

\[
\Delta P_i, \Delta M_i = f((F_i, F_t))
\]

where \( \Delta P_i = \{ \Delta p_i, k \}, \Delta M_i = \{ \Delta m_i, k \} \), \( f \) denotes the prediction function, and \( [\cdot, \cdot] \) represents concatenation.

As shown in Fig.3, the input features first pass through a \( 3 \times 3 \) convolution. Then, convolutions with different dilation rates are applied to extract multi-level features and halve the number of channels simultaneously. After that, multi-level features are concatenated together and pass through a \( 1 \times 1 \) convolution layer for feature fusion and restoring the number of channels to the same as that of the input.

As we stack several deformable convolutions, features in more shallow layers need larger receptive fields to obtain more accurate parameters while the deep features only require relatively small ones to achieve precise alignment. Thus, we adjust the number of multi-dilation convolutions and the dilation rates according to the position of the deformable convolution. The deeper, the less.

Starting from the first deformable convolution, the number of dilation convolutions in the DMDCU of the subsequent deformable convolutions will decrease by two as the network progress. For example, in our experiments, we set the number of deformable convolutions \( n^{dc} \) to be 4, and the number of dilation convolutions \( n^{dil} \) in the DMDCU of the first deformable convolution to be 8. Then, \( n^{dil} \) in the next three deformable convolutions are 6, 4, and 2, respectively. Meanwhile, the dilation rates range from 1 to 8, 1 to 6, 1 to 4 and 1 to 2, correspondingly.

In practice, if \( n^{dc} \) exceeds 4, we keep \( n^{dil} \) in the first several deformable convolutions to be 8 until the third from the last deformable convolution, and then sequentially reduce \( n^{dil} \).
by two. In other words, the amount of dilation convolutions is at most 8 to prevent too much computation. Ablation study on the case of \( n^{dc} = 5 \) is shown in Section 5. Using the novel DMDCUs, we promote the ability of predicting more accurate deformable sampling parameters. Moreover, the decreasing design avoids the addition of too many parameters and reduces redundant calculations.

**C. LOCAL FUSION MODULE**

After the alignment module, details from different time period need to be effectively fused so that features of neighboring frames can be benefit for better reconstruction of the reference frame. In order to maintain the temporal consistency and impose tight coherence between adjacent frames, we design a local fusion module (LFM) which consists of several local fusion blocks (LFBs).

![Figure 4. An example of seven input features passing through an LFB in LFM.](image)

Both the input and the output of an LFB are a series of features corresponding to the video frames. An example of seven input features passing through an LFB is depicted in Fig.4. When the features pass through the \( b \)-th LFB in turn, the current input features \( L_{t}^{b-1} \) will be fused with the last output features of previous frame \( L_{t-1}^{b-1} \) in the LFB. Note that if the \( L_{t}^{b-1} \) is the first input, \( L_{t-1}^{b-1} \) is the copy of \( L_{t}^{b-1} \). The output has two copies, one as the result of the current input and the other to be sent to the next frame as part of the input features.

![Figure 5. The detailed structure of a local fusion block (LFB).](image)

The detailed structure of an LFB is shown in Fig.5. For each current input \( L_{t}^{b-1} \), it is first concatenated with the previous output features \( L_{t-1}^{b} \) and they pass through a 3 \( \times \) 3 convolution for feature fusion. Then, the fused features are concatenated with \( L_{t}^{b-1} \) again. In this way, the current features contain information from previous fused features and itself. After that, they go through a 1 \( \times \) 1 convolution for integration and channel reduction. Two residual blocks are used to extract features and an element-wise addition is applied for residual learning. At last, the output \( L_{t}^{b} \) comprises the detail information of previous frames and temporal coherence. The LFB which handles features of one frame can be defined as:

\[
L_{t}^{\text{temp}} = \text{Conv}_{3}(L_{t-1}^{b}, L_{t}^{b-1})
\]

\[
L_{t}^{b} = \text{Net}_{\text{RB}}(\text{Conv}_{1}(L_{t}^{\text{temp}}, L_{t}^{b-1}))) + L_{t}^{b-1}
\]

where \( L_{t}^{\text{temp}} \) is the intermediate feature maps, \([\cdot, \cdot]\) denotes concatenation and \( \text{Net}_{\text{RB}} \) represents the two residual blocks. Finally, \( \{L_{t-N}^{b}, \ldots, L_{t}^{b}, \ldots, L_{t+N}^{b}\} \) represents the output of the \( b \)-th LFB.

The previous output features contain most information of the former frame due to the concatenation and residual connection in LFB. Besides, the characteristics used to maintain temporal consistency are also included implicitly. Therefore, the reuse of previous output features fuses the knowledge of two adjacent frames locally and makes the information continuously flow from the former frame to the latter. The stacked LFBs constantly strengthen the connection between adjacent frames and ensure the continuity of video sequences.

**D. GLOBAL FUSION MODULE**

Although LFM solves the problem of feature combination from a local perspective, the latter features cannot be fully combined with the former features, which results in a waste of potential space characteristics. Therefore, we present a global fusion module (GFM) which consists of stacked global fusion blocks (GFBs). GFB exquisitely exchanges features among different video frames and fuses them together by taking advantage of a video shuffle strategy. This not only fills the defect of LFM, but also allows the information in different frames with long temporal interval to be fully covered.

![Figure 6. Video shuffle and inverted video shuffle strategy.](image)

Video shuffle (VS), firstly introduced by Ma et al. [32], is utilized for exploring video representation which is essentially a channel switching tactic. As depicted in Fig.6, VS is a bijective transformation. The features of each frame are divided equally into the same number of groups as that of the input frames. For example, if there are \( T \) input frames,
from each of which is extracted a feature map of $C$ channels, then each feature map will be divided into $T$ groups, each consisting of $\lfloor C/T \rfloor$ channels. Note that we use a $3 \times 3$ convolution layer in GFB to make sure the output channel divided by $T$ is integer and then we perform grouping. After that, the groups with the same group id are rearranged to form new feature maps, each of which consists of a certain amount of information belonging to various periods. The inverted process is to reverse the groups that belong to the same frame back to the original feature maps.

In order to take advantage of features from multiple frames and fuse the abundant features comprehensively, we enhance the capability of VS by using multi-kernel convolution. Our proposed GFB is shown in Fig.7. After the forward shuffle, each feature map will pass through a $1 \times 1$ convolution layer to reduce the channels. Then, convolutions with various kernel sizes are responsible for exploring multi-level features. Note that to reduce parameters and promote efficiency, we have utilized two $3 \times 3$ convolutions to implement an equivalent $5 \times 5$ convolution. Afterwards, features are concatenated together and sent into a $3 \times 3$ convolution for feature fusion. The outputs of inverted video shuffle are sent into another $3 \times 3$ convolution followed by a residual skip connection. The application of multi-kernel convolution can discover more abundant details and integrate them from diverse period effectively, compared to the common VS strategy.

**FIGURE 7.** The detailed structure of a global fusion block (GFB). The input features exchange information globally through a video shuffle strategy and fuse with each other by multi-kernel convolution. The inverted video shuffle restores the channels into the original sequence.

Within each GFB, the exchange of features from various frames fuses plentiful information together. Moreover, correlations among frames are learned implicitly. The whole GFM ensures that spacial characteristics in every input frames can be fully utilized by each other, in case of missing important details.

After the two feature fusion branches, we now obtain two kinds of features of consecutive frames. One from the LFM maintains high temporal consistency and the other from the GFM combines overall characteristics. As shown in Fig.1, for two kinds of features which represent each frame, we merge them together through a simple $1 \times 1$ convolution, which focuses on the relationship of diverse patterns.

**E. RECONSTRUCTION MODULE**

In this module, the fusion features are aggregated together and pass through the advanced reconstruction blocks for restoring high quality reference frame. Here, we utilized the sophisticated residual channel attention blocks (RCABs). Note that it can be replaced by more effective blocks if necessary.

The channel attention mechanism essentially obtains a series of weights coefficient to various channels and multiplies them with feature maps, so that the network can bring different emphasises on the features and enhance the ability of learning representation. Details about channel attention mechanism for SR reconstruction can be found in [39]. In the end, the restored features pass through a pixel shuffle layer to produce high-resolution clear reference frame in RGB format.

**IV. EXPERIMENTS**

**A. DATASET AND TRAINING DETAILS**

1) **DATASET**

Since there is no unified training dataset for VSR, we adopt a widely used dataset Vimeo-90K collected by Xue et al. [26] for network training. 64612 training samples with a fixed resolution of $448 \times 256$ are included. They contain not only manifold, but also intricate motion and scenes. To generate the LR input, MATLAB imresize function, which blurs the input with cubic filters and then downsamples them through bicubic interpolation, is applied for $4 \times$ downscaling.

2) **TRAINING DETAILS**

Through the whole network, the filter number is set to 64 and the kernel size is set to $3 \times 3$ for the convolution if not specified. The shallow feature extractor applies five residual blocks. Four deformable convolutions are utilized for frame alignment. The number of dilation convolutions in each DMDCU is $[8, 6, 4, 2]$ respectively. Both LFM and GFM have 10 internal blocks for feature fusion. At the end, 20 RCABs are used for SR reconstruction.

It is observed that each training sample in Vimeo-90K contains seven frames. Thus, our full model chooses seven continuous frames as inputs, which are cropped randomly into fixed size patches during the end-to-end training process. The learning rate is set to $10^{-4}$ for the first 50 epochs, and declines by half every 20 epochs. Besides, the Adam optimizer [44] is chosen with $\beta_1 = 0.9$, $\beta_2 = 0.999$, and $\epsilon = 10^{-8}$. We adopt simple $L_1$ loss to minimize the distance between the restored reference frame and the ground truth HR. After convergence, we fine-tune the model slightly by $L_2$ loss for better performance. All experiments were conducted on two NVIDIA RTX 2080ti GPUs using PyTorch 1.0.1 [45].

**B. COMPARISON TO STATE-OF-THE-ART METHODS**

Several state-of-the-art methods, including RCAN [39], TOFlow [26], FRVSR [20], DUF [31], RBPN [29] and TDAN [30], are used to compare with our proposed LGFN. We either directly utilize the results they provided or download their pretrained model to predict the SR frames. All quantitative results are evaluated on Y channel (i.e., luminance) from the transformed YCbCr space with both the...
TABLE 1. Quantitative results of state-of-the-art SR algorithms on Vid4 for 4×. Bold type indicates the best and underline indicates the second best performance (PSNR/SSIM). During inference, the first and last frames are not participated. Only DUF needs to crop borders and eight pixels near image boundary are eliminated.

| Clip Name | Flow Magnitude | Beorhst (1 Frame) | RCAN (1 Frame) | TOFlow (7 Frames) | PRVSR (recurrent) | DUF (7 Frames) | RBPN (7 Frames) | TDAN (3 Frames) | LGFN (Ours) (7 Frames) |
|-----------|---------------|-------------------|----------------|-------------------|-------------------|----------------|----------------|----------------|---------------------|
| Calendar  | 1.14          | 20.59 / 0.5720    | 22.31 / 0.7248 | 22.44 / 0.7290    | -                 | 22.07 / 0.8123 | 23.95 / 0.8076 | 23.20 / 0.7689 | 24.11 / 0.8313     |
| City      | 1.63          | 23.17 / 0.6024    | 26.07 / 0.6958 | 26.75 / 0.7368    | -                 | 28.32 / 0.8333 | 27.74 / 0.8051 | 27.18 / 0.7716 | 27.35 / 0.8099     |
| Foliage   | 1.18          | 23.47 / 0.5666    | 24.69 / 0.6629 | 25.24 / 0.7065    | -                 | 26.41 / 0.7713 | 26.21 / 0.7579 | 25.64 / 0.7224 | 26.53 / 0.7526     |
| Walk      | 1.44          | 26.11 / 0.7977    | 28.64 / 0.8718 | 29.03 / 0.8777    | -                 | 30.63 / 0.9164 | 30.70 / 0.9111 | 29.80 / 0.8940 | 30.93 / 0.9137     |
| Average   | 1.42          | 23.79 / 0.6347    | 25.53 / 0.7383 | 25.86 / 0.7625    | 26.69 / 0.822     | 27.36 / 0.8328 | 27.15 / 0.8204 | 26.46 / 0.7907 | 26.31 / 0.8248     |

PSNR and the SSIM [46] metrics. Note that the first and last two frames are discarded for a stable comparison. In order to prevent the severe boundary effects, we crop eight pixels near image boundary only for the DUF results as suggested in [29].

Experiments are conducted on three public testing datasets: Vid4 [47], SPMCS [24] and Vimeo-90K-T [26]. Average flow magnitude (pixel/frame) is presented following [29]. Commonly reported Vid4 consists of four video sequences including calendar, city, foliage and walk. Note that the inter-frame motion is restricted and some artifacts exist on a few regions of the ground truth. On the other hand, SPMCS contains more varied and clean scenes with abundant texture information which brings higher requirements for the models. In addition, Vimeo-90K-T, with large number of video frames, comprises manifold circumstances and objects. Various light transformation and multiple motion types rise great challenges for the restoration ability, which allows us to distinguish the strengths and weaknesses of different methods.

Table 1 presents the results on Vid4. Our model attains comparative performance and achieves the best PSNR on calendar and walk clips. More visual comparisons are depicted in Fig.8. It can be seen that our method restores the date in the calendar without deformation while others show up certain blur or undesired artifacts. Moreover, LGFN portrays the correct white boundary of the building in the city clip while others reduce the clarity.

Since SPMCS has 32 testing sequences, we show the quantitative outcomes in Table 2 and Table 3 respectively. The former are the average results of all clips while the latter correlate with several individual clips. Apparently, our method achieves the best performance and surpasses both SISR and VSR method. The abundant texture information in the dataset requires the network to learn precise high frequency features, which confirms that our approach possesses such kind of capability. Qualitative comparisons are shown in Fig.9. In the first row, the areas with dense textures were effectively restored with our LGFN. Meanwhile, the distant and small street lamp in the fifth row can well distinguish different methods with stronger abilities. Our approach is the only one that produces intact details and correct pattern while others suffer from varying degrees of indistinctness.

For a more comprehensive comparison, we divided the Vimeo-90K-T into three subsets according to the motion velocities as [29]. Quantitative results are shown in Table 4. With the motion velocity increases, LGFN surpasses RBPN by 0.16 dB, 0.25 dB and 0.34 dB respectively, which indicates that our model can handle video frames under different situations. High improvement in fast motion velocity especially reflects the superiority of our model which takes fully advantage of temporal information. On the contrary, others with limited representation ability exhibit poor outcomes. Visual evaluations are presented in Fig.10. When the color of the texture is similar to the surrounding environment, such as row one, our model is the exclusive one which can restore similar pattern with the ground truth. Other examples give evidences that LGFN can recover detail information accurately in various scenarios.

C. SUPER-RESOLVING REAL-WORLD VIDEO SEQUENCES

Since most of the existing approaches are trained on simulated datasets, it is challenging to restore satisfied real-world video sequences which contain camera shake and complex image formation methods. To make the results of LGFN more convincing, we conduct experiments on real-world dataset provided by Tao et al. [24] for evaluation. The dataset includes 30 video clips which are obtained through both SLR camera and mobile phone. Quantitative comparisons are shown in Table 5. Here, we utilize two reference-free image quality assessment indices NIQE [48] and SSEQ [49] to measure the performance of the methods. Both indices range from 0 (best) to 100 (worst). It can be noticed that our LGFN achieves the best performance under NIQE, which further demonstrates the effectiveness of our model. Moreover, visual results are presented in Fig.11. Our model can not only restore the correct text information but also the detailed textures, while other methods could suffer from blur or distortion.

D. MODEL SIZE AND RUNNING TIME ANALYSES

We compare the model size and running time of several methods. All pretrained models are downloaded from the public official repositories. We conduct the entire testing experiments on the machine with 3.6GHz Inter i7 CPU (16G RAM) and single NVIDIA RTX 2080ti GPU. In addition, the corresponding dependencies are strictly followed. The input frames for testing is with size 112 × 64. As shown in Table 6, our model achieves the best result and maintains the model size into a moderate level. In the meantime, LGFN runs faster than the optical-flow based approach TOFlow and dynamic kernel prediction method DUF, which verifies the efficiency of our network. The application of dilation convolutions in DMDCU may be a little time consuming, which explains why RBPN (without dilation convolutions) runs slightly
faster even it has more parameters. However, it is worthy to sacrifice a small amount of time to promote performance significantly.

V. ABLATION STUDY

In order to explore the exact contribution of each module in our framework, we perform more experiments to demonstrate
their effectiveness. We first remove every module individually to observe how much influence these components exert on the network. The results are presented in Table 7. It can be observed that eliminating the alignment module including deformable convolutions and the DMDCUs damages the restoration ability significantly. Excessive differences between frames make it difficult for other parts of the network to learn effective information. The model can only utilize very limited features and even introduce inaccurate artifacts from other frames. Moreover, deletion of the fusion module also leads to reduction of the performance. It proves that the two kinds of fusion strategy can take advantage of abundant features locally and globally, which indicates the ascendancy of our proposed method. As for the reconstruction module, normal residual blocks displace RCABs to recover video frames. The evaluation metrics decrease slightly compared to other components which means the improvement of performance is mainly caused by other two modules rather than the channel attention mechanism.

Since the alignment module plays an indispensable role, we further investigate the functionality from the following two aspects: one is the decreased multi-dilation convolution unit (DMDCU), and the other is the deformable convolution (DC).

We first explore the impact of DMDCU by altering the number of dilation convolutions in each unit. The results are shown in Table 8. Following [30], we set the number of DC to four. The first column represents the number of dilation convolutions in each DMDCU. For example, [8, 6, 4, 2] indicates that the alignment module has four DCs and the number of dilation convolutions in each DMDCU is 8, 6, 4 and 2 respectively from shallow to deep DC. Also, the dilation rates range from 1 to 8, 1 to 6, 1 to 4 and 1 to 2 correspondingly. 0 means that in this DC we apply ordinary convolution rather than the DMDCU to predict the parameters. Through comparing [8, 6, 4, 2] and [8, 8, 8, 8] group, the usage of more dilation convolutions just promotes the performance slightly. The decreased strategy can avoid many redundant calculations so as to reduce the cost of space and time. Meanwhile, applying DMDCU in the first two DCs can further improve performance and reducing DMDCU can lower the restoration effect.

Moreover, we adjust the number of DCs to see if the performance is subject to fluctuations. As shown in Table 9, we gradually increase the number of DCs, accompanying with the corresponding DMDCUs. It can be observed that the restoration ability improves more significantly with the increase of DCs, compared to the impact of the number of dilation convolutions in each DMDCU shown in Table 8.

| Table 4. Quantitative results of state-of-the-art SR algorithms on Vimeo-90K-T for 4 ×. Bold type indicates the best and underline indicates the second best performance (PSNR/SSIM). |
|-----------------|---------|---------|---------|---------|
| Method          | Slow    | Medium  | Past    | Average |
| Bicubic         | 29.34 / 0.8330 | 31.29 / 0.8708 | 34.07 / 0.9050 | 31.32 / 0.8684 |
| RCAN            | 32.93 / 0.9002 | 35.35 / 0.9268 | 38.47 / 0.9466 | 35.34 / 0.9249 |
| TOPflow         | 32.15 / 0.8900 | 33.61 / 0.9238 | 37.70 / 0.9440 | 34.84 / 0.9209 |
| DUF             | 33.41 / 0.9110 | 36.71 / 0.9466 | 38.87 / 0.9510 | 36.37 / 0.9386 |
| RBPN            | 34.26 / 0.9222 | 37.39 / 0.9494 | 40.16 / 0.9611 | 37.18 / 0.9456 |
| TDAN            | 33.20 / 0.9067 | 36.07 / 0.9367 | 38.68 / 0.9492 | 35.68 / 0.9325 |
| LGF (ours)      | 34.42 / 0.9237 | 37.64 / 0.9512 | 40.50 / 0.9634 | 37.42 / 0.9473 |
| # of clips      | 1616    | 4983    | 1225    | 7824    |
| Flow Mag.       | 0.6     | 2.5     | 8.3     | 3.0     |
This demonstrates the pivotal role of alignment module which enables the features of the neighboring frames to be effectively utilized by the following modules. Besides, the performance improves more prominently when adds the first two DCs, which indicates the shallow features need to be aligned more accurately. After we increase the number of DC into 5, the PSNR only promotes 0.03 db, which means four DCs are a reasonable trade-off between performance and computational cost.
Next, more experiments are conducted on the two-branch structure to reveal the effect of combining various fusion strategies. We only keep one branch at a time to study the impact of a single branch on the entire network. For a fairer comparison, we additionally compensate the parameter number to be close to the complete model for a fairer comparison. The results are presented in Table 10. Notice that applying a single fusion strategy can not bring more benefits to the model. In the meantime, simply adding their number of blocks improves performance slightly but introduces an unworthy increase in computational cost. Therefore, our framework takes advantage of both feature fusion tactics and unites them for a better reconstruction output. Compared with a single fusion branch, the integration method promotes at least 0.11 dB on PSNR.

In order to explain the different concerns of the two branches more intuitively, we visualize the feature maps of the reference frames after different fusion strategies and the integration results of two branches. As depicted in Fig. 12, the local fusion features supplement detail textures while the global fusion features describe the overall outline structure. The integrated features of two branches using 1 × 1 convolution present clear and accurate high frequency information. The reason for this phenomenon is as follows. The local fusion module combines features of adjacent frames and maintains the temporal consistency, which makes the branch pay more attention to the difference between high frequency features.

---

**FIGURE 12.** Feature maps visualization of reference frame on SPMCS. (a) represents the original reference frames. (b) and (c) show the feature maps after local fusion module and global fusion module, respectively. (d) shows the integration results of two branches after 1 × 1 convolution.

**TABLE 7.** Ablation study of proposed network on Vimeo-90K-T for 4×. We only remove one module at a time to explore the performance of each module. **Bold type** indicates the best results.

| Alignment module | Fusion module | Reconstruction module | PSNR | SSIM  |
|------------------|---------------|-----------------------|------|-------|
| ✓                | ✓             | ✓                     | 36.64| 0.9399|
| ✓                | ✓             | ✓                     | 36.95| 0.9430|
| ✓                | ✓             | ✓                     | 37.29| 0.9464|
| ✓                | ✓             | ✓                     | 37.42| 0.9475|

**TABLE 8.** Ablation study of number of dilation convolutions in DMDCU on Vimeo-90K-T for 4×. Four deformable convolutions (DCs) are set as default and we only adjust the number of dilation convolutions. **Bold type** indicates the best results.

| Number of dilation convolutions in each DMDCU | PSNR | SSIM  |
|-----------------------------------------------|------|-------|
| [0, 0, 0, 0]                                  | 37.35| 0.9458|
| [8, 0, 0, 0]                                  | 37.29| 0.9462|
| [8, 6, 0, 0]                                  | 37.36| 0.9468|
| [8, 6, 4, 0]                                  | 37.40| 0.9471|
| [8, 6, 4, 2]                                  | 37.42| 0.9473|
| [8, 8, 8]                                    | 37.44| 0.9474|

**TABLE 9.** Ablation study of number of deformable convolutions (DCs) on Vimeo-90K-T for 4×. The amount of DC gradually increases as well as the DMDCU. **Bold type** indicates the best results.

| Number of deformable convolutions | Number of dilation convolutions in each DMDCU | PSNR | SSIM  |
|----------------------------------|-----------------------------------------------|------|-------|
| 1                                | [8]                                           | 36.88| 0.9425|
| 2                                | [8, 6]                                        | 37.17| 0.9453|
| 3                                | [8, 6, 4]                                     | 37.34| 0.9467|
| 4                                | [8, 6, 4, 2]                                  | 37.42| 0.9473|
| 5                                | [8, 8, 8, 4, 2]                               | 37.45| 0.9475|
regions of adjacent frames. In the meantime, since the global fusion branch exchanges information through video shuffle globally, abundant features from all input frames are effectively utilized. In the end, two kinds of fusion features are integrated together and the outputs show splendid outcomes which contain both clear textures and outlines.

Finally, we study the impact of different amount of frames for restoration. This factor determines how much initial information can be provided. Since the VSR task focuses on the effective usage of spatial and temporal information, less number of features will undoubtedly weaken the capability of the network. As shown in Table 11, three, five and seven LR frames are utilized as inputs to train three models respectively. We can see that the evaluation metrics become better with increasing number of input frames as more abundant features can be explored.

**TABLE 11.** Ablation study of different number of input frames on Vimeo-90K-4 for 4x. *Bold type indicates the best results.*

| Input | 3 frames | 5 frames | 7 frames |
|-------|----------|----------|----------|
| PSNR / SSIM | 36.33 / 0.9411 | 37.13 / 0.9445 | 37.42 / 0.9473 |

**VI. CONCLUSION AND FUTURE WORK**

In this paper, we propose a novel local-global fusion network (LGFN) for high quality video super-resolution. In order to handle the implicit motion compensation for input consecutive frames, LGFN adopts improved deformable convolutions with decreased multi-dilation convolution units for effective frame alignment. A structure with two branches is introduced for local and global feature fusion. The local fusion module focuses on combining adjacent features and maintaining the temporal consistency while the global fusion module figures out how to fully take advantage of all input frames. Extensive experimental results on benchmark datasets demonstrate that our model can achieve comparative performance with state-of-the-art methods and restore a variety of video frames efficiently.

Although improved deformable convolutions with decreased multi-dilation convolution units perform implicit alignment well, the aliasing effect could still exist in the aligned neighboring features, which means that the alignment may be incorrect in some cases. One possible reason is that the alignment process is performed in low-resolution scale, which indicates that less information could be utilized and it results in adverse impact on alignment. Besides, there is no reasonable supervision approach to guarantee the alignment accuracy. To overcome such drawbacks, a valuable direction is to apply the attention mechanism to discover the area that has not been aligned precisely. The misaligned positions can be corrected appropriately through suppression operation. Spatial attention and its variants are potential choices. Moreover, other prior knowledge can be introduced to enhance the ability of alignment. For example, segmentation maps or depth information can be used to guide the alignment process to obtain more accurate results. It would be our great interest to explore possible solutions.

**REFERENCES**

[1] S. Y. Kim, J. Oh, and M. Kim, “FISR: Deep joint frame interpolation and super-resolution with a multi-scale temporal loss,” in *Proc. AAAI Conf. Artif. Intell.*, 2020, pp. 11278–11286.

[2] H. Liu, Y. Gu, T. Wang, and S. Li, “Satellite video super-resolution based on adaptively spatiotemporal neighbors and nonlocal similarity regularization,” *IEEE Trans. Geosci. Remote Sens.*, early access, Apr. 28, 2020, doi: 10.1109/TGRS.2020.2987400.

[3] S. Y. Kim, J. Oh, and M. Kim, “Deep SR-ITM: Joint learning of super-resolution and inverse tone-mapping for 4K UHD HDR applications,” in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2019, pp. 3116–3125.

[4] H. Seibel, S. Goldenstein, and A. Rocha, “Eyes on the target: Super-resolution and license-plate recognition in low-quality surveillance videos,” *IEEE Access*, vol. 5, pp. 20020–20035, 2017.

[5] Y. Lee, J. Yun, Y. Hong, J. Lee, and M. Jeon, “Accurate license plate recognition and super-resolution using a generative adversarial networks on traffic surveillance video,” in *Proc. IEEE Int. Conf. Consum. Electron.-Asia*, Jun. 2018, pp. 1–4.

[6] D. S. Jeon, S.-H. Baek, I. Choi, and M. H. Kim, “Enhancing the spatial resolution of stereo images using a parallax prior,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 1721–1730.

[7] L. Wang, Y. Wang, Z. Liang, Z. Lin, J. Yang, W. An, and Y. Guo, “Learning parallax attention for stereo image super-resolution,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2019, pp. 12250–12259.

[8] W. Song, S. Choi, S. Jeong, and K. Sohn, “Stereoscopic image super-resolution with stereo consistent feature,” in *Proc. AAAI Conf. Artif. Intell.*, 2020, pp. 12031–12038.

[9] C. Duan and N. Xiao, “Parallax-based spatial and channel attention for stereo image super-resolution,” *IEEE Access*, vol. 7, pp. 183672–183679, 2019.

[10] S. Zhang, Y. Lin, and H. Sheng, “Residual networks for light field image super-resolution,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2019, pp. 11046–11055.

[11] N. Meng, X. Wu, J. Liu, and E. Y. Lam, “High-order residual network for light field super-resolution,” in *Proc. AAAI Conf. Artif. Intell.*, 2020, pp. 11757–11764.

[12] J. Jin, J. Hou, J. Chen, and S. Kwong, “Light field spatial super-resolution via deep combinatorial geometry embedding and structural consistency regularization,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2020, pp. 2260–2269.

[13] D. C. Garcia, C. Doreca, and R. L. de Queiroz, “Super resolution for multiview images using depth information,” *IEEE Trans. Circuits Syst. Video Technol.*, vol. 22, no. 9, pp. 1249–1256, Sep. 2012.

[14] E. Faramarzi, D. Rajan, and M. P. Christensen, “Unified blind method for multi-image super-resolution and Single/Multi-image blur deconvolution,” *IEEE Trans. Image Process.*, vol. 22, no. 6, pp. 2101–2114, Jun. 2013.

[15] S. Huang, J. Sun, Y. Yang, Y. Fang, and P. Lin, “Multi-frame super-resolution reconstruction based on gradient vector flow hybrid field,” *IEEE Access*, vol. 5, pp. 21669–21683, 2017.

[16] B. Wronski, I. Garcia-Dorado, M. Ernst, D. Kelly, M. Krainin, C.-K. Liang, M. Levoy, and P. Milanfar, “Handheld multi-frame super-resolution,” *ACM Trans. Graph.*, vol. 38, no. 4, pp. 1–18, Jul. 2019.

[17] W. Wang, C. Ren, X. He, H. Chen, and L. Qing, “Video super-resolution via residual learning,” *IEEE Access*, vol. 6, pp. 23767–23777, 2018.

[18] K.-W. Hung, C. Qiu, and J. Jiang, “Video super-resolution via deep global-aware network,” *IEEE Access*, vol. 7, pp. 74711–74720, 2019.

[19] S. Li, X. Chen, S. Z. Zhou, W. Pan, and Y. Wang, “SAR analysis for real video super-resolution improvement,” *IEEE Access*, vol. 8, pp. 54816–54832, 2020.

[20] M. M. S. Najidi, R. Vemulapalli, and M. Brown, “Frame-recurrent video super-resolution,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 6626–6634.

[21] B. Yan, C. Lin, and W. Tan, “Frame and feature-context video super-resolution,” in *Proc. AAAI Conf. Artif. Intell.*, vol. 33, 2019, pp. 5597–5604.

[22] X. Zhu, Z. Li, X.-Y. Zhang, C. Li, Y. Liu, and Z. Xue, “Residual invertible spatio-temporal network for video super-resolution,” in *Proc. AAAI Conf. Artif. Intell.*, vol. 33, 2019, pp. 5981–5988.

[23] D. Fuoli, S. Gu, and R. Timofte, “Efficient video super-resolution through recurrent latent space propagation,” in *Proc. IEEE Int. Conf. Comput. Vis. Workshop*, Oct. 2019, pp. 3476–3485.
[24] X. Tao, H. Gao, R. Liao, J. Wang, and J. Jia, “Detail-revealing deep video super-resolution,” in Proc. IEEE Int. Conf. Comput. Vis., Oct. 2017, pp. 4472–4480.

[25] Z. Wang, P. Yi, K. Jiang, J. Jiang, Z. Han, T. Lu, and J. Ma, “Multi-memory convolutional neural network for video super-resolution,” IEEE Trans. Image Process., vol. 28, no. 5, pp. 2530–2544, May 2019.

[26] T. Xue, B. Chen, J. Wu, D. Wei, and W. T. Freeman, “Video enhancement with task-oriented flow,” Int. J. Comput. Vis., pp. 1–20, 2017.

[27] F. Li, H. Bai, and Y. Zhao, “Learning a deep dual attention network for video super-resolution,” IEEE Trans. Image Process., vol. 29, pp. 4474–4488, 2020.

[28] X. Liu, L. Kong, Y. Zhou, J. Zhao, and J. Chen, “End-to-end trainable video super-resolution based on a new mechanism for implicit motion estimation and compensation,” in Proc. IEEE Winter Conf. Appl. Comput. Vis., Mar. 2020, pp. 2416–2425.

[29] Y. Jo, S. W. Oh, J. Kang, and S. J. Kim, “Deep video super-resolution network using dynamic upsampling filters without explicit motion compensation,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., Jun. 2018, pp. 3224–3232.

[30] P. Ma, Y. Zhou, Y. Lu, and W. Zhang, “Learning efficient video representation with video shuffle networks,” 2019, arXiv:1911.11319. [Online]. Available: http://arxiv.org/abs/1911.11319

[31] C. Dong, C. C. Loy, K. He, and X. Tang, “Learning a deep convolutional network for image super-resolution,” in Proc. Eur. Conf. Comput. Vis., 2014, pp. 184–199.

[32] C. Dong, C. C. Loy, and X. Tang, “Accelerating the super-resolution convolutional neural network,” in Proc. Eur. Conf. Comput. Vis., 2016, pp. 391–407.

[33] W. Shi, J. Caballero, F. Huszar, J. Totz, A. P. Aitken, R. Bishop, D. Rueckert, and Z. Wang, “Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., Jun. 2016, pp. 1874–1883.

[34] T. Tong, G. Li, X. Liu, and Q. Gao, “Image super-resolution using dense skip connections,” in Proc. IEEE Int. Conf. Comput. Vis., Oct. 2017, pp. 4799–4807.

[35] Y. Zhang, Y. Tian, Y. Kong, B. Zhong, and Y. Fu, “Residual dense network for image super-resolution,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., Jun. 2018, pp. 2472–2481.

[36] T. Dai, J. Cai, Y. Zhang, S.-T. Xia, and L. Zhang, “Second-order attention network for single image super-resolution,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., Jun. 2019, p. 11.

[37] Y. Zhang, K. Li, K. Li, L. Wang, B. Zhong, and Y. Fu, “Image super-resolution using very deep residual channel attention networks,” in Proc. Eur. Conf. Comput. Vis., 2018, pp. 286–301.

[38] P. Yi, Z. Wang, K. Jiang, Z. Shao, and J. Ma, “Multi-temporal ultra dense memory network for video super-resolution,” IEEE Trans. Circuits Syst. Video Technol., vol. 30, no. 8, pp. 2503–2516, Aug. 2020.

[39] L. Wang, Y. Guo, L. Liu, Z. Lin, X. Deng, and W. An, “Deep video super-resolution using HR optical flow estimation,” IEEE Trans. Image Process., vol. 29, pp. 4323–4336, 2020.

[40] J. Dai, H. Qi, Y. Xiong, Y. Li, G. Zhang, H. Hu, and Y. Wei, “Deformable convolutional networks,” in Proc. IEEE Int. Conf. Comput. Vis., Oct. 2017, pp. 764–773.

[41] X. Zhu, H. Hu, S. Lin, and J. Dai, “Deformable ConvNets v2: More deformable, better results,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., Jun. 2019, pp. 9308–9316.

[42] D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” 2014, arXiv:1412.6980. [Online]. Available: http://arxiv.org/abs/1412.6980

[43] A. Paszke et al., “Pytorch: An imperative style, high-performance deep learning library,” in Proc. Adv. Neural Inf. Process. Syst., 2019, pp. 8026–8037.

[44] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, “Image quality assessment: From error visibility to structural similarity,” IEEE Trans. Image Process., vol. 13, no. 4, pp. 600–612, Apr. 2004.

[45] C. Liu and D. Sun, “On Bayesian adaptive video super-resolution,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 36, no. 2, pp. 346–360, Feb. 2014.

[46] A. Mittal, R. Soundararajan, and A. C. Bovik, “Making a ‘completely blind’ image quality analyzer,” IEEE Signal Process. Lett., vol. 20, no. 3, pp. 209–212, Nov. 2012.

[47] L. Liu, B. Liu, H. Huang, and A. C. Bovik, “No-reference image quality assessment based on spatial and spectral entropies,” Signal Process., Image Commun., vol. 29, no. 8, pp. 856–863, Sep. 2014.

DEWEI SU received the B.S. degree in software engineering from the South China University of Technology, China, in 2019, where he is currently pursuing the M.S. degree in software engineering. His current research interests include video super-resolution and deep learning.

HUA WANG received the B.S. degree from the East China University of Science and Technology, in 2016. He is currently pursuing the M.S. degree in software engineering with the South China University of Technology, China. His current research interests include image/video restoration and deep learning.

LONGCUN JIN (Member, IEEE) is currently an Associate Professor with the School of Software Engineering, South China University of Technology, China. His research projects include the National Natural Science Foundation of China, the Doctoral Program Fund of Ministry of Education, Science and Technology Program of Guangdong Province, the Open Research Fund of State Key Laboratory, the Science and Technology Program of Guangzhou City, and the Key Projects of Basic Scientific Research Operating Expenses of Central Universities of Ministry of Education. His research interests include computer vision and artificial intelligence. He is a member of ACM, AAAI, AAAS, and SIAM. He has received the Baosteel Award, the Hong Kong Scholars Award, and the Pearl River Science and Technology Star.

XINYI PENG was a Former Vice President and the Executive Vice President of the South China University of Technology, China. He is currently a Full Professor with the School of Software Engineering, South China University of Technology. He has presided over and participated in more than 20 projects, with the project funds of over nine million. Among them, nine projects above the provincial level have passed the acceptance inspection. His current research interests include artificial intelligence and data mining.

XINYI PENG received the Ph.D. degree from the Institute of Automation, Chinese Academy of Sciences, in 1994. He is currently a Senior Lecturer with the School of Computer Science, Cardiff University, U.K. His main research interests include computer vision, computer graphics, pattern recognition, and artificial intelligence.