Multi-Stage Feature Fusion Network for Video Super-Resolution

Huihui Song, Wenjie Xu, Dong Liu, Bo Liu, Qingshan Liu, *Senior Member, IEEE*, and Dimitris N. Metaxas, *Fellow, IEEE*

***Abstract*— Video super-resolution (VSR) is to restore a photo-realistic high-resolution (HR) frame from both its corresponding low-resolution (LR) frame (reference frame) and multiple neighboring frames (supporting frames). An important step in VSR is to fuse the feature of the reference frame with the features of the supporting frames. The major issue with existing VSR methods is that the fusion is conducted in a one-stage manner, and the fused feature may deviate greatly from the visual information in the original LR reference frame. In this paper, we propose an end-to-end Multi-Stage Feature Fusion Network that fuses the temporally aligned features of the supporting frames and the spatial feature of the original reference frame at different stages of a feed-forward neural network architecture. In our network, the Temporal Alignment Branch is designed as an inter-frame temporal alignment module used to mitigate the misalignment between the supporting frames and the reference frame. Specifically, we apply the multi-scale dilated deformable convolution as the basic operation to generate temporally aligned features of the supporting frames. Afterwards, the Modulative Feature Fusion Branch, the other branch of our network accepts the temporally aligned feature map as a conditional input and modulates the feature of the reference frame at different stages of the branch backbone. This enables the feature of the reference frame to be referenced at each stage of the feature fusion process, leading to an enhanced feature from LR to HR. Experimental results on several benchmark datasets demonstrate that our proposed method can achieve state-of-the-art performance on VSR task.**

***Index Terms*— Video super-resolution, single image super-resolution, deep learning, deformable convolution, feature fusion.**

I. Introduction

VSR aims to recover an HR video frame from its corresponding LR frame [1]–[5]. Different from Single Image super-resolution (SISR) that only needs to take into account one single LR image [6]–[11], VSR exploits both the input LR frame (reference frame) and multiple neighboring LR frames surrounding it (supporting frames). HR video frames contain fine-grained visual details and would bring visual-pleasing experience to the video consumers. As a result, effective VSR techniques are increasingly demanding in various real-world applications such as video surveillance, high-definition television (HDTV), video post-production, to name a few [12]–[14].

Given the nature of VSR, modeling the temporal relations between the LR reference frame and the neighboring LR supporting frames is critical for the success of the VSR performance [2], [16], [18], [19]. Due to the camera or object motion commonly observed in video, the LR reference frame and each of the LR supporting frames may not align well, and a vital task of VSR is to accurately align them such that the temporal redundancy among neighboring frames can be fully exploited. The state-of-the-art methods [16], [17] tackle this task by employing the deformable convolutions (DConvs) to adaptively align the reference frame and each of the supporting frames at the feature level. In specific, the features extracted from both the reference frame and each supporting frame are leveraged to predict the offsets of sampling the convolution kernels, through which the learned dynamic kernels are applied on the features of supporting features to perform temporal alignment. Based on the aligned features of the supporting frames and the reference frame, a fusion operation is conducted to aggregate all the features into a holistic feature representation of the entire temporal sequence, which is then used as input for reconstructing the HR frame.

Despite promising performance, the major issue with these existing methods is that the feature fusion operation does not take into account the discrepancy between the fused feature and the visual information in the original LR reference frame. Such discrepancy may arise from the imperfect feature alignment or severe blurring caused by intensive motions in the video, which, if not handle properly, may be further amplified in the upstream HR frame reconstruction operation, leading to degraded VSR performance as shown in Figure 1. Ideally, the feature fusion should be performed in a progressive manner in which the original LR reference frame can intervene the fusion over multiple stages of the fusion process so that the fused feature can faithfully preserve the visual information in the reference frame and be used to accurately reconstruct an HR frame (see ours in Figure 1).

Motivated by the issue above, we propose an end-to-end *Multi-Stage Feature Fusion Network* that fuses the temporally aligned features of the supporting frames and the spatial

Manuscript received July 16, 2020; revised December 8, 2020 and January 7, 2021; accepted January 27, 2021. Date of publication February 9, 2021; date of current version February 16, 2021. This work was supported in part by the National Natural Science Foundation of China under Grant 61872189 and Grant 61825601, in part by the Natural Science Foundation of Jiangsu Province under Grant BK20191397, and in part by the Postgraduate Research and Practice Innovation Program of Jiangsu Province under Grant KYCX20_0968. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Jiaying Liu. *(Corresponding author: Huihui Song.)*

Huihui Song, Wenjie Xu, and Qingshan Liu are with the Jiangsu Key Laboratory of Big Data Analysis Technology (B-DAT), Nanjing University of Information Science and Technology, Nanjing 210044, China, and also with the Jiangsu Collaborative Innovation Center on Atmospheric Environment and Equipment Technology (CICAEET), Nanjing University of Information Science and Technology, Nanjing 210044, China (e-mail: songhuihui@nuist.edu.cn).

Dong Liu is with Netflix Inc., Los Gatos, CA 95032 USA.

Bo Liu is with JD Finance America Corporation, Mountain View, CA 94089 USA.

Dimitris N. Metaxas is with the Department of Computer Science, Rutgers University, Piscataway, NJ 08854 USA.

Digital Object Identifier 10.1109/TIP.2021.3056868

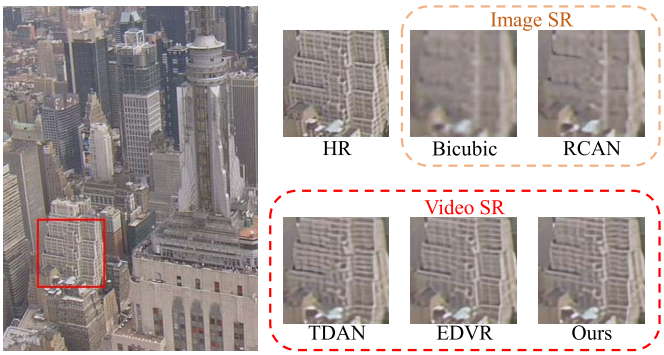


Fig. 1. The ×4 SR results of our method compared with existing state-of-the-art SISR and VSR algorithms, including RCAN [15], TDAN [16] and EDVR [17]. We can observe that our method can restore more accurate image structure details than the state-of-the-arts.

feature of the original reference frame at different stages of a feed-forward neural network architecture. Figure 2 illustrates the difference between our multi-stage fusion method and the existing one-stage fusion method [1], [16], [17], [20]. In our network, the *Temporal Alignment Branch* is designed at inter-frame temporal alignment module that can be used to mitigate the misalignment between the supporting frames and the reference frame at the feature level. Given a reference frame and a supporting frame, we employ the DConv [21] commonly applied in VSR [16], [17] to align their features. To effectively explore the context information in video frames, we propose to use multi-scale dilated convolution as the basic operation to learn the offsets of sampling the convolution kernels so that object/scene of different scales can be better aligned across frames (c.f. Section III-B). Our experimental results show that such a simple multi-scale dilated deformable alignment module outperforms the existing pyramid, cascading and deformable (PCD) alignment [17] (see Table V, a significant gain of PSNR score = 0.14 dB on Vid4 dataset). The output of the temporal alignment branch is a feature map of all supporting frames.

The *Modulative Feature Fusion Branch* accepts the aforementioned temporally aligned feature map as a conditional input to multiple *Modulative Residual Fusion Blocks* (MRFBs) cascaded as the branch backbone. Each MRFB accepts as input both the spatially transformed feature map of the reference frame and the temporally aligned conditional input feature map. The temporally aligned conditional feature map is used to learn a pair of modulation parameters that can be applied as an affine transformation on the spatial feature map of the reference frame. Applying the modulation parameters on the spatial feature map, we end up with a temporally modulated spatial feature map, which preserves the visual information in the original reference (via skip connection) and can be passed to the next MRFB for the next stage of fusion. We also feed the output of the last MRFB back to the first one to effectively combine the low-level and the high-level fusion feature maps. The advantage of our method is that the visual information of the reference frame can flow over different stages of the feature fusion process, allowing the spatial and temporal features to make deep mutual interactions guided by the VSR learning goal. We will demonstrate experimentally that the proposed method can achieve favorable performance over state-of-the-arts when evaluated over the benchmark VSR datasets.

We highlight the main contributions of our work:

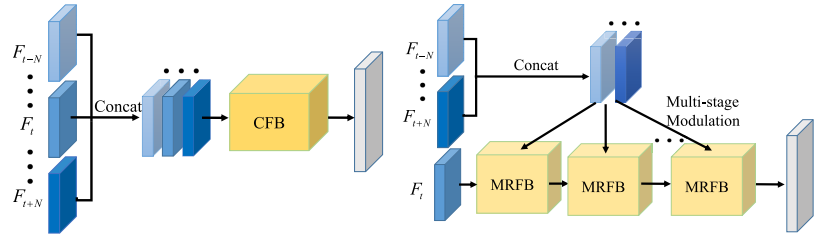
- A novel VSR feature fusion method allowing spatial and temporal features to be aggregated at different stages of a network backbone.
- A multi-scale deformable alignment module to align frames at the feature level.
- State-of-the-art performance on benchmark VSR datasets.

II. Related Work

a) Video Super-resolution: Depending on how the multiple input frames are being processed, existing VSR methods can be divided into two categories. The *first* considers video as time-series data, and processes the frames one by one through Recurrent Neural Networks (RNNs) [2], [18], [22]. Huang *et al*. [18] use a bidirectional recurrent architecture to model the inter-frame relations in video without explicitly leveraging the motion compensation for frame alignment. Tao *et al*. [2] approach VSR within the framework of convolutional long short-term memory (ConvLSTM) [23] and propose a sub-pixel motion compensation operation to project LR frames onto HR image space while achieving resolution enhancement. Sajjadi *et al*. [22] propose a Frame-Recurrent VSR method (FRVSR), which uses the super-resolved HR frame of the previous frame as input to super-resolve the current frame. The *second* category directly takes multiple LR frames as supportive evidence and generates a single HR output frame from them. Kappeler *et al*. [24] warp the previous frame and the next frame onto the current frame using optical flow and then concatenate the three frames as input, feeding them to a CNN to produce the output HR frame. Caballero *et al*. [1] follow the similar idea and introduce the first end-to-end VSR network which jointly trains flow estimation and spatio-temporal sub-networks. This inspires a recent line of research on solving VSR through end-to-end learning [2], [3], [5], [19]. Our method also falls into this category.

A critical operation in end-to-end VSR is feature fusion. Most existing methods either use convolutions to perform early fusion on all frames [1], [16], [24], [25], and only achieve limited success due to the ignorance of the underlying visual information on each frame. To treat the frames and the spatial locations in them differently, Wang *et al*. [17] propose a Temporal and Spatial Attention (TSA) feature fusion module to merge multiple frames whose features are aligned via the PCD alignment module inspired by the TDAN method proposed in [16]. However, the problem with these existing methods is that they just perform feature fusion in a one-stage fashion, which may not faithfully preserve the original visual information in the reference frame. Our method relieves this issue and shows superiority over the existing methods.

b) Network Conditioning: Our work is inspired by previous studies on feature normalization. The widely used technique is Batch normalization (BN), which can ease the network training by normalizing feature statistics [26]. Later, Conditional Normalization (CN) is introduced to replace the



(a) One-stage fusion (b) Multi-stage fusion

Fig. 2. Different fusion methods in VSR. (a) One-stage fusion method concatenates the features of multiple frames as input to a Convolutional Fusion Block (CFB) [1], [16], [17], [20]. (b) Our multi-stage fusion method accepts supporting frames as condition and fuses them into reference frame through a cascade of Modulative Residual Fusion Blocks (MRFBs) (see Section III-C).

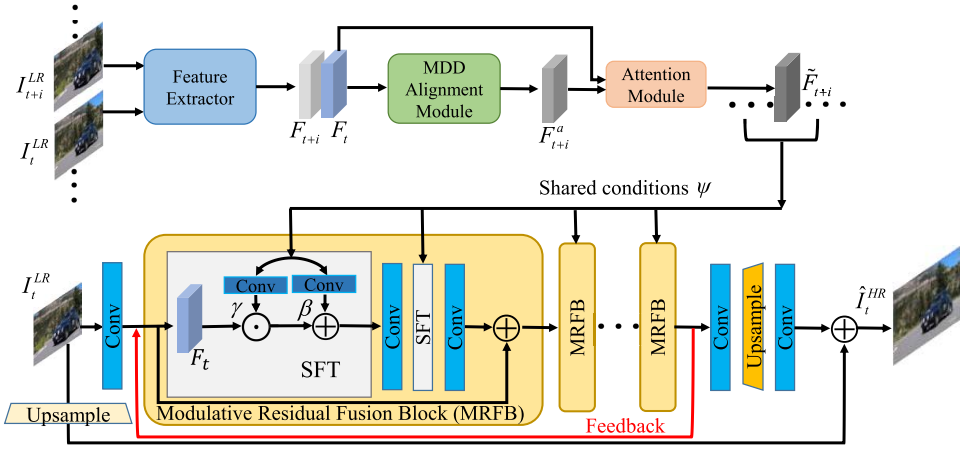


Fig. 3. Overview of the proposed MSFFN for VSR. Top branch is the temporal alignment sub-network f_{TAN}. Here, we only demonstrate it with one supporting frame. Bottom branch is the modulative feature fusion sub-network f_{MFFN}, among which the temporally aligned features from f_{TAN} are accepted as the shared conditions to progressively modulate the intermediate feature maps of f_{MFFN} and adapt it to encode the rich temporal information that is essential for accurate VSR.

parameters of affine transformation in BN with a learned function of some conditions. CN has been widely applied in image style transfer [27]–[29], visual question answering [30] and visual reasoning [31], [32]. Perez *et al.* [32] develop a Feature-wise Linear Modulation (FiLM) layer, and show that the affine transformation in CN needs not be placed after normalization, indicating that the features can be directly modulated. Wang *et al.* [33] realize that FiLM cannot be applicable to SR due to its incapability of preserving spatial information of image, and therefore propose a Spatial Feature Transform (SFT) layer to modulate the image feature with spatial condition information (in the form of image segmentation map). Different from these existing methods, our method accepts the temporally aligned features of the supporting frames as condition, and uses them to learn multiple sets of modulation parameters used to transform spatial features of the reference frame at different stages of the network backbone. To our best knowledge, this is the first work applying conditional features for multi-stage feature fusion.

c) Deformable Convolution: Dai *et al.* [34] first propose DConv, in which additional offsets are learned to allow the network to obtain information away from its regular local neighborhood, improving the capability of regular convolutions. DConvs are widely used in various vision tasks such as video object detection [35], action recognition [36] and semantic segmentation [34]. Tian *et al.* [15] first introduce DConvs into VSR and use them to align the input frames at the feature level without explicit motion estimation or image warping. Inspired by TDAN [15], Wang *et al.* [17] devise a PCD alignment module, extending it to a pyramid architecture and aligning features from coarse to fine. We also apply DConvs to align frames at the feature level, but the key difference is that the offsets in our DConvs are learned through multi-scale dilated convolutions that can well preserve object boundary details [37] and therefore brings in useful multi-scale context information into the alignment procedure.

III. Methodology

A. Overview of the Proposed Framework

Given a sequence of $2N+1$ LR frames $\mathcal{I} = \{I_{t+i}^{LR}\}_{i=-N}^{N}$, where the center frame I_t^{LR} denotes the reference frame and the others are supporting frames, our goal is to reconstruct the HR reference frame $\hat{I}_t^{HR}$ that best approaches to the ground-truth frame I_t^{HR}. To this end, we design a Multi-Stage Feature Fusion Network (MSFFN) f_{MSFFN} to predict $\hat{I}_t^{HR}$ as

$$\hat{I}_t^{HR} = f_{MSFFN}(\mathcal{I}). \tag{1}$$

Figure 3 shows the framework of f_{MSFFN}, including two sub-networks: a Temporal Alignment Network (TAN) f_{TAN} and a Modulative Feature Fusion Network (MFFN) f_{MFFN}. The f_{TAN} accepts the reference frame I_t^{LR} and one supporting frame I_{t+i}^{LR} as input and estimates the aligned features $\tilde{F}_{t+i}$ of the corresponding supporting frame as

$$\tilde{F}_{t+i} = f_{TAN}(I_t^{LR}, I_{t+i}^{LR}). \tag{2}$$

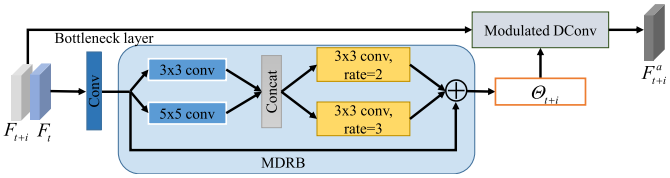


Fig. 4. Illustration of the MDD alignment module. The features F_{t+i} and F_t are combined by a convolutional layer and fed into the MDRB to predict the sampling parameters Θ_{t+i}. The MDRB consists of two parallel $\{3 \times 3, 5 \times 5\}$ kernels and two 3×3 kernels with different dilation rates $\{2, 3\}$ that can extract multi-scale features with enlarged receptive field. Finally, the modulated DConv with the parameters Θ_{t+i} accepts the features F_{t+i} as input, generating the corresponding aligned features F^a_{t+i}.

Afterwards, all the aligned features of the supporting frames are concatenated as

$$\psi = [\tilde{F}_{t-N}, \ldots, \tilde{F}_{t-1}, \tilde{F}_{t+1}, \ldots, \tilde{F}_{t+N}]. \tag{3}$$

Inspired by the spatial feature transform (SFT) in [33], we take ψ as the shared SFT conditions, which encode the useful temporally aligned feature information to progressively fuse with the multi-stage features of the I^{LR}_t, yielding the predicted HR reference frame $\hat{I}^{HR}_t$ as

$$\hat{I}^{HR}_t = f_{MFFN}(I^{LR}_t | \psi). \tag{4}$$

B. Temporal Alignment Network

Given the LR frames I^{LR}_{t+i} and I^{LR}_t, the f_{TAN} temporally aligns I^{LR}_{t+i} with I^{LR}_t by looking at spatio-temporal neighborhoods of pixels, which can avoid explicit motion compensation. The f_{TAN} includes three modules: feature extraction module, Multi-scale Dilated Deformable (MDD) alignment module and attention module.

1) Feature Extraction Module: The feature extraction module is consist of one convolutional layer and 5 residual blocks [38] with ReLU activation function. We extract the features F_{t+i} and F_t from I^{LR}_{t+i} and I^{LR}_t with the shared feature extraction module, and feed them into the MDD alignment module.

2) MDD Alignment Module: Figure 4 shows the architecture of the proposed MDD alignment module. The input F_{t+i} and F_t are concatenated and fed into a 3×3 bottleneck layer to reduce the channels of the feature maps. Recently, TDAN [16] directly feeds the features maps into a convolutional layer to predict the sampling parameters Θ of the DConv [34]. However, the feature maps have a restricted receptive field that may make it unable to see the corresponding pixel pairs between frames $t + i$ and t, especially when suffering from large and complex motion. Afterwards, EDVR [17] extends the sampling parameter estimation in TDAN by exploring multi-scale feature maps generated by strided convolution filtering. However, EDVR misses detailed information of object boundaries due to the convolutions with striding operations on the feature maps, leading to an inaccurate estimation. To address the issues above, we design a multi-scale dilated residual block (MDRB) f_{MDRB} that can not only effectively enlarge receptive field to sense large pixel motions between frames, but also capture multi-scale contextual information that can well preserve object boundary details with the help of dilated convolutions [37].

As illustrated in Figure 4, the f_{MDRB} first stacks two 3×3 and 5×5 convolutional kernels in parallel to extract multi-scale features. Then, the features are fed into two 3×3 kernels with different dilation rates $= 2, 3$ that are beneficial to enlarge receptive field with less gridding effect [39]. Such a simple design can effectively enlarge receptive field with much lower computational cost than the PCD alignment module in EDVR [17]. Hence, the MDRB helps to exploit the temporal dependency of pixels between frames even when suffering from complex and large motions, generating accurate sampling parameters Θ as

$$\Theta_{t+i} = f_{MDRB}([F_{t+i}, F_t]), \tag{5}$$

where $\Theta_{t+i} = \{\Delta p_k, \Delta m_k, k = 1, \ldots, K\}$. Here, we adopt the modulated DConv [21]: $\Delta p_k \in \mathcal{R}^2$ and $\Delta m_k \in \mathcal{R}$ denote the learning offset and the modulation scalar therein respectively.

With the learned sampling parameters Θ_{t+i}, the aligned features F^a_{t+i} at each position $p \in \mathcal{R}^2$ can be obtained by

$$F^a_{t+i}(p) = \sum_{k=1}^{K} w_k \cdot F_{t+i}(p + p_k + \Delta p_k) \cdot \Delta m_k, \tag{6}$$

where p_k represents the k-th sampling grid of K sampling locations and w_k denotes the corresponding weight. For instance, a 3×3 kernel is defined with $K = 9$ and $p_k \in \{(-1, -1), (-1, 0), \ldots, (1, 1)\}$. As $p + p_k + \Delta p_k$ may be fractional, we use the bilinear interpolation as in [34]. In (6), the DConv is operated on the irregular positions with dynamic weights, which can adaptively sample the input features. Conceptually, the adaptive sampling in (6) depends on the pixel motions as its sampling parameters Θ_{t+i} (5) are predicted by looking at wide spatio-temporal neighborhoods of pixels, enabling us to well handle large and complex motions.

3) Attention Module: Although the MDD alignment module has the potential to align F_{t+i} with F_t, some misalignments are easy to arise due to occlusion, blurry regions and parallax problems, making the aligned features at different spatial positions not equally informative, thereby leading to a large discrepancy to the reference frame on some features. To address this issue, we design a spatial attention mask M to weight the F^a_{t+i} in (6)

$$\tilde{F}_{t+i} = F^a_{t+i} \odot M_{t+i}, \tag{7}$$

where M_{t+i} measures the pixel-wise similarity between F_t and F^a_{t+i}, which is defined as

$$M_{t+i}(x, y) = \frac{1}{1 + \exp\left(|F_t(x, y, :) - F^a_{t+i}(x, y, :)|^1_1\right)}, \tag{8}$$

where we leverage the L1 distance $|\cdot|^1_1$ to pay more attention to the features at highly-confident locations $(x, y) \in \mathcal{R}^2$.

C. Modulative Feature Fusion Network

Existing state-of-the-art methods for VSR [16], [17], [20] often first fuse the features of the reference frame and the aligned supporting frames by concatenation and then feed them into a reconstruction network to generate an HR output. However, this simple one-stage fusion strategy has two limitations: First, the aligned supporting frames and the reference frame have a quantity of similar patterns at feature level,

so simply concatenating them will introduce a large amount of redundancy into the reconstruction network, leading to expensive computational cost. Second, the fusion only happens at the beginning layer, and hence the complementary temporal information from the supporting frames will be gradually weakened with the layers of the deep network being more and more deeper.

To address the above issues, as illustrated by the bottom branch of Figure 3, we propose the f_{MFFN} that cascades a set of MRFBs plugged at different depths of the branch backbone. We employ the high-level architecture of SRResNet [8] as the branch backbone. Each MRFB contains an SFT layer [33] that takes the temporally aligned features ψ in (3) as the shared conditions to modulate its input feature maps F_t from the reference frame. The SFT layer outputs affine transformation of F_t conditioned on ψ by a scaling and shifting operation as

$$f_{SFT}(F_t|\psi) = \gamma \odot F_t + \beta, \tag{9}$$

where γ and β is the parameters for scaling and shifting, $\odot$ denotes pixel-wise multiplication. The transformation parameters γ and β can be generated by feeding ψ into several convolutional layers with different weights. We inject SFT layers after all convolutional layers in each MRFB, which consistently enhances the visual information of the reference frame with the aligned temporal information in the multi-stage fusion process. Finally, we feed the high-level features learned from the last MRFB back to the input layer of the first one through a feedback skip-connection. This feedback mechanism refines the low-level features with high-level information, and the refined features are passed through the modulative feature fusion network, facilitating to learn a complex nonlinear mapping from the LR to the HR image space without extra parameters.

IV. Experimental Analysis

A. Benchmark Datasets and Evaluation Metrics

We use Septuplet, a subset of Vimeo-90K dataset collected by Xue *et al.* [41] to train our model. We follow the standard training and testing dataset splits provided by this public dataset. The training set contains 64, 612 videos. Each video has 7 frames with a fixed resolution of 448×256. To generate LR videos, we resize the original HR frames to $4\times$ resolution as 112×64 with MATLAB *imresize* function, which first blurs the input frames using cubic filters and then downsamples them using bicubic interpolation. We augment the training dataset by random horizontal flipping and 90° frame rotation.

After training the model on Septuple, we test it on three benchmark datasets, including Vid4 [4], SPMC-11 [2], and Vimeo-90K-T [41]. Following the popular SR model evaluation approaches [3], [17], we adopt PSNR and SSIM [42] calculated on luminance (Y) channel of transformed YCbCr space to evaluate the proposed model by comparing its performance with a variety of state-of-the-art methods.

B. Implementation Details

We only leverage the Charbonnier penalty function [43] as the loss, defined by $\mathcal{L} = \sqrt{\|\hat{I}_t^{HR} - I_t^{HR}\|_2^2 + \varepsilon^2}$, where ε is set to $1e-3$. The f_{MFFN} contains 16 MFRBs. The channel size in each layer is set to 80 for final comparison (termed as *Ours*) and 64 for ablation study (termed as *Ours-S*). The network takes 7 consecutive frames (*i.e.,* $N = 3$) as inputs unless otherwise specified. In each training batch, 16 LR RGB patches with the size of 64×64 are extracted as input. We initiate the network parameters using the method in [44] and update them by the Adam optimizer [45] with $\beta_1 = 0.9$, $\beta_2 = 0.999$. The initial learning rate is set to $4e-4$. All the experiments have been conducted on NVIDIA RTX 2080Ti GPUs using PyTorch 1.0.

C. Comparison With State-of-the-Arts

We compare our method with state-of-the-art SISR and VSR methods. For SISR methods, we compare with Bicubic, RCAN [15] and DBPN [40]. The compared VSR methods include VESPCN [1], $B_{123}+T$ [19], SPMC [2], TOFlow [41], FRVSR [22], DUF [20], RBPN [3] and EDVR [17].[1] To obtain the video PSNR and SSIM evaluation, we calculate the PSNR and SSIM scores of every frame excluding the first and last two in the video, and then average the values as video quality measurement. For DUF method, eight pixels near image boundary are cropped due to its severe boundary effects.

1) Quantitative Evaluation: Tables I−III list the quantitative results achieved on the Vid4 dataset, the SPMC-11 dataset and the Vimeo-90K-T dataset.

a) Evaluation on Vid4 Dataset: Vid4 is a widely-used benchmark dataset which contains four video sequences: Calendar, City, Foliage and Walk. We also provide the average flow magnitude (pixel/frame) [3], which means that the videos in this set have limited inter-frame variance, and visual artifacts also exist on its ground-truth frames. Table I lists the PSNR and SSIM results of all considered methods on the Vid4 testing set. It is shown that our model outperforms other methods except for DUF in terms of average PSNR and SSIM metrics. However, our model is better than DUF on Calendar and Walk clips which proves the superiority of our model. Specifically, the PSNR value on Walk clip achieved by our framework is higher than DUF by 0.22 dB. That is because, large motion offsets can be well predicted by our MDD alignment module and therefore more reliable spatial details and more accurate temporal information can be aggregated into the features of the current frame, which is very helpful to reconstruct the HR frame accurately.

b) Evaluation on SPMC-11 Dataset: We next test the proposed method on SPMC-11 dataset, which consists of 11 high-quality video clips with various motions and diverse scenes. We present PSNR and SSIM results of all methods on SPMC-11 dataset in Table II. Our model achieves the best results in terms of both PSNR and SSIM compared with both SISR and VSR approaches. In particular, our model has better performance of 0.70 dB and 0.08 dB than DUF and RBPN in terms of PSNR, respectively. From Table II we can see that our method achieves the highest PSNR and SSIM values on 8 and 7 out of 11 testing videos and the average value of all testing video clips, compared with other SISR and VSR

[1]For EDVR method, we set the model size same as ours for fair comparison.

TABLE I

QUANTITATIVE COMPARISONS ON VID4 FOR 4×. RED INDICATES THE BEST AND BLUE INDICATES THE SECOND BEST PERFORMANCE (PSNR/SSIM). '#PARAM." REPRESENTS THE NUMBER OF PARAMETERS. FLOPS IS CALCULATED ON AN HR IMAGE OF SIZE 720 × 480. RUNTIME IS THE TOTAL TIME TESTED ON VID4 AND IS AVERAGED OVER 20 RUNS

Algorithm	#Param.	FLOPs	Runtime	Clip Name/Flow Magnitude				Average
				Calendar	City	Foliage	Walk	
				1.14	1.63	1.48	1.44	
Bicubic	-	-	-	20.39/0.5720	25.16/0.6028	23.47/0.5666	26.10/0.7974	23.78/0.6347
DBPN [40]	10M	11.6T	34.9s	22.27/0.7178	25.84/0.6835	24.70/0.6615	28.65/0.8706	25.37/0.7334
RCAN [15]	16M	0.38T	134.5s	22.33/0.7254	26.10/0.6960	24.74/0.6647	28.65/0.8719	25.46/0.7395
VESPCN [1]	0.88M	-	23.6s		-	-	-	25.35/0.7557
$B_{123}+T$ [20]	-	-	-	21.66/0.7040	26.45/0.7200	24.95/0.6980	28.26/0.8590	25.34/0.7450
SPMC [2]	-	-	-	22.16/0.7465	27.00/0.7573	25.43/0.7208	28.91/0.8761	25.88/0.7752
TOFlow [41]	1.4M	0.81T	59.4s	22.47/0.7318	26.78/0.7403	25.27/0.7092	29.05/0.8790	25.89/0.7651
FRVSR [22]	5.1M	0.14T	-	-	-	-	-	26.69/0.8220
DUF [18]	6.8M	0.62T	92.6s	24.04/0.8110	28.27/0.8313	26.41/0.7709	30.60/0.9141	27.33/0.8318
RBPN [3]	12.8M	9.30T	98.8s	23.93/0.8030	27.64/0.8020	26.27/0.7570	30.65/0.9110	27.12/0.8180
EDVR [17]	5.5M	0.91T	30.5s	23.82/0.8038	27.66/0.7977	26.06/0.7523	30.52/0.9077	27.02/0.8153
Ours	8.5M	1.02T	51.2s	24.06/0.8117	27.81/0.8050	26.22/0.7582	30.82/0.9123	27.23/0.8218

TABLE II

QUANTITATIVE COMPARISON ON SPMC-11 FOR 4×. RED INDICATES THE BEST AND BLUE INDICATES THE SECOND BEST PERFORMANCE (PSNR/SSIM). "F.M." REPRESENTS THE FLOW MAGNITUDE

Clip Name	F.M.	Bicubic	RCAN [15]	SPMC [2]	TOFlow [41]	DUF [18]	RBPN [3]	Ours
car_05	6.21	27.75/0.7825	29.86/0.8484	32.19/0.9103	30.10/0.8626	30.79/0.8707	31.95/0.9021	32.02/0.9033
hdclub_003	0.70	19.42/0.4863	20.41/0.6096	21.04/0.6752	20.86/0.6253	22.05/0.7438	21.91/0.7257	22.09/0.7331
hitachi_isee5	3.01	19.61/0.5938	23.71/0.8369	23.76/0.8296	22.88/0.8044	25.77/0.8929	26.30/0.9049	26.48/0.9065
hk004_001	0.49	28.54/0.8003	31.68/0.8631	32.13/0.8788	30.89/0.8654	32.98/0.8988	33.38/0.9016	33.42/0.9027
HKVTG_004	0.11	27.46/0.6831	28.81/0.7649	28.78/0.7665	28.49/0.7487	29.16/0.7860	29.51/0.7979	29.51/0.7974
jvc_009	1.24	25.40/0.7558	28.31/08717	28.24/0.8642	27.85/0.8542	29.18/0.8961	30.06/0.9105	30.34/0.9153
NYVTG_006	0.10	28.45/0.8014	31.01/0.8859	31.41/0.8903	30.12/0.8603	32.30/0.9090	33.22/0.9231	33.05/0.9225
PRVTG_012	0.12	25.63/0.7136	26.56/0.7806	27.02/0.7970	26.62/0.7788	27.39/0.8166	27.60/0.8242	27.71/0.8271
RMVTG_011	0.18	23.96/0.6573	26.02/0.7569	26.43/0.7766	25.89/0.7500	27.56/0.8133	27.63/0.8170	27.75/0.8192
veni3_011	0.36	29.47/0.8979	34.58/0.9629	34.77/0.9576	32.85/0.9536	34.63/0.9677	36.61/0.9735	36.22/0.9735
veni5_015	0.36	27.41/0.8483	31.04/0.9262	31.58/0.9246	30.03/0.9118	31.88/0.9371	32.37/0.9409	32.83/0.9450
Average	1.17	25.73/0.7291	28.36/0.8279	28.85/0.8428	27.87/0.8220	29.43/0.8664	30.05/0.8747	30.13/0.8769

TABLE III

QUANTITATIVE COMPARISONS ON VIMEO-90K-T FOR 4×. RED INDICATES THE BEST AND BLUE INDICATES THE SECOND BEST PERFORMANCE (PSNR/SSIM)

Algorithm	Bicubic	RCAN [15]	TOFlow [41]	DUF [18]	RBPN [3]	EDVR [17]	Ours
Average	31.32/0.8684	35.35/0.9251	34.83/0.9220	36.37/0.9387	37.07/0.9435	36.98/0.9435	37.33/0.9467

baseline methods. In comparison to Vid4, SPMC-11 contains more high-frequency information with higher resolution which requires the superb recovery abilities of algorithms. Moreover, our model exceeds the optical flow based methods SPMC [2] and TOFlow [41], demonstrating the effectiveness of our MDD alignment approach.

c) Evaluation on Vimeo-90K-T Dataset: Vimeo-90K-T is a much larger and diverse data benchmark testing set containing 7,824 videos. Table III shows the quantitative PSNR and SSIM results of all methods on this dataset. The results are obtained by first calculating the PSNR and SSIM scores of each video, then averaging the values of all videos. It can be observed that RBPN achieves the best performance among all baseline methods while our method attains higher PSNR and SSIM scores compared with it. Our method gets higher PSNR scores by 0.26 dB and 0.35 dB than RBPN and EDVR respectively. Since the number of video clips of Vimeo-90K-T is much larger than Vid4 and SPMC-11, as suggested in RBPN [3], the video clips can be classified into three different groups (*e.g.* slow, medium and fast) according to the motion velocity. That is to say, this dataset requires the algorithms which are able to solve different degrees of motion problems between video frames. Table III reflects that our method enables to take full advantage of information among multi-frames effectively and address various motion problems flexibly.

2) Qualitative Results: Figure 5 illustrates the qualitative results on two scenarios of the Vid4 dataset. It can be observed from the zoom-in regions that our framework recovers finer and more reliable details. In the frame example from "Calendar" video, our method recovers the clearest numbers "31" than others. It is noticeable that the ground truth of HR

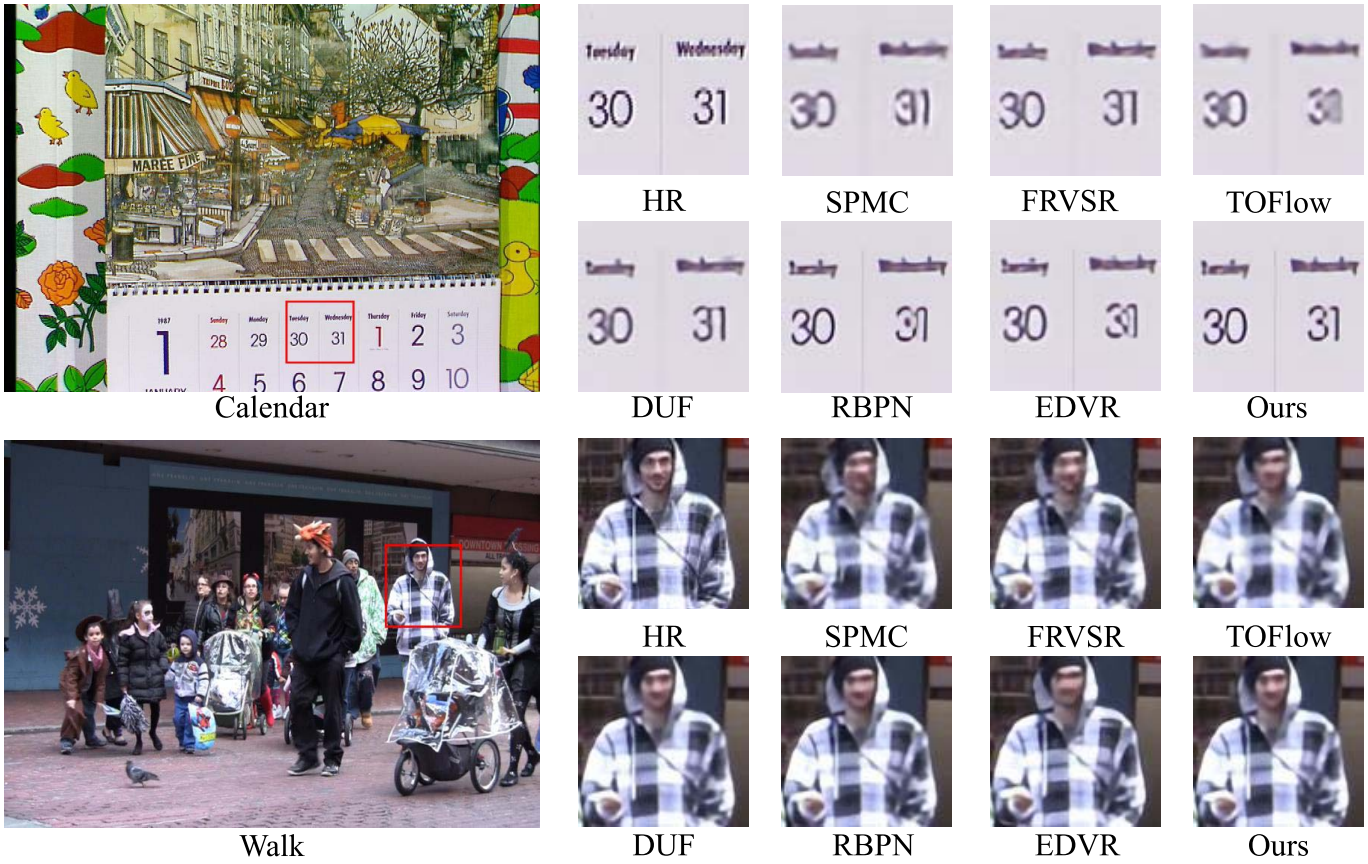


Fig. 5. Qualitative results on Vid4 for 4× scaling factor.

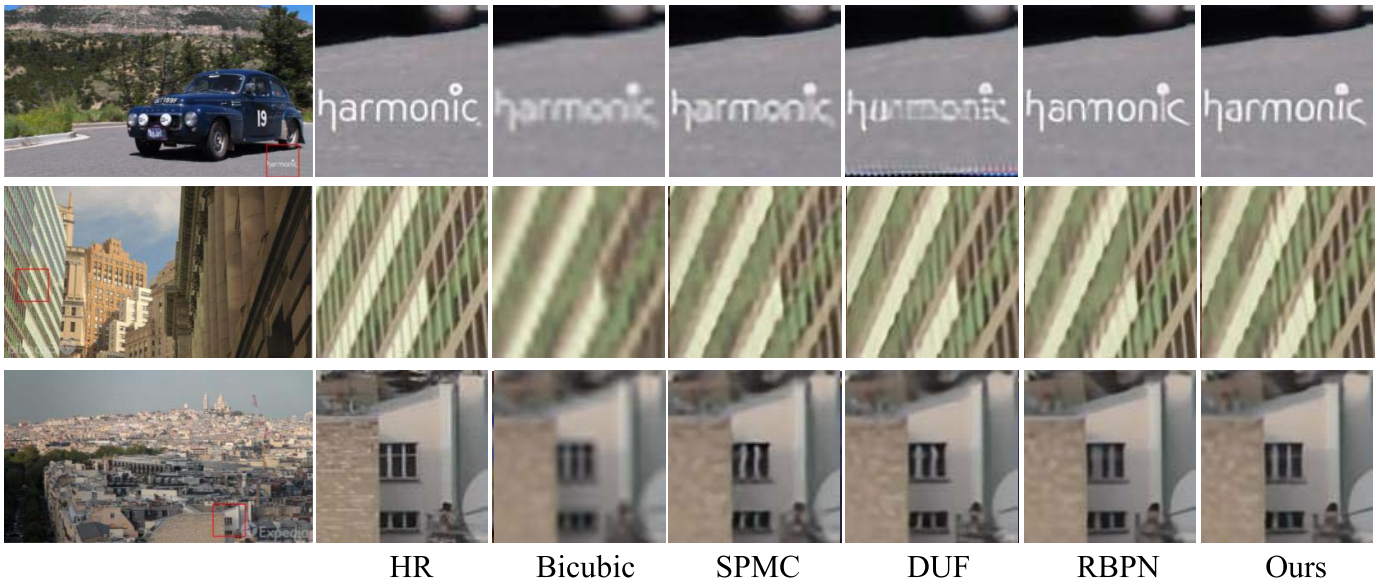


Fig. 6. Qualitative results on SPMC-11 for 4× scaling factor.

frame from "Walk" video has artifact. However, as the walk clip displays that most previous methods blur the rope and clothing together, only our model can clearly distinguish these two parts and restore the rope pattern closest to the ground truth frame. The qualitative comparisons on SPMC-11 dataset are depicted in Figure 6. It is observable that previous methods tend to produce severe artifacts. For example, in the example shown in the first row, although the watermark "harmonic" is fixed across different frames, as the background changes, other algorithms tend to be affected by temporal information while ours is able to distill pure information and recover the true word. In the second and the third rows, the compared methods fail to recover high-quality edge information of the windows. The output of our model has higher-quality edge information.

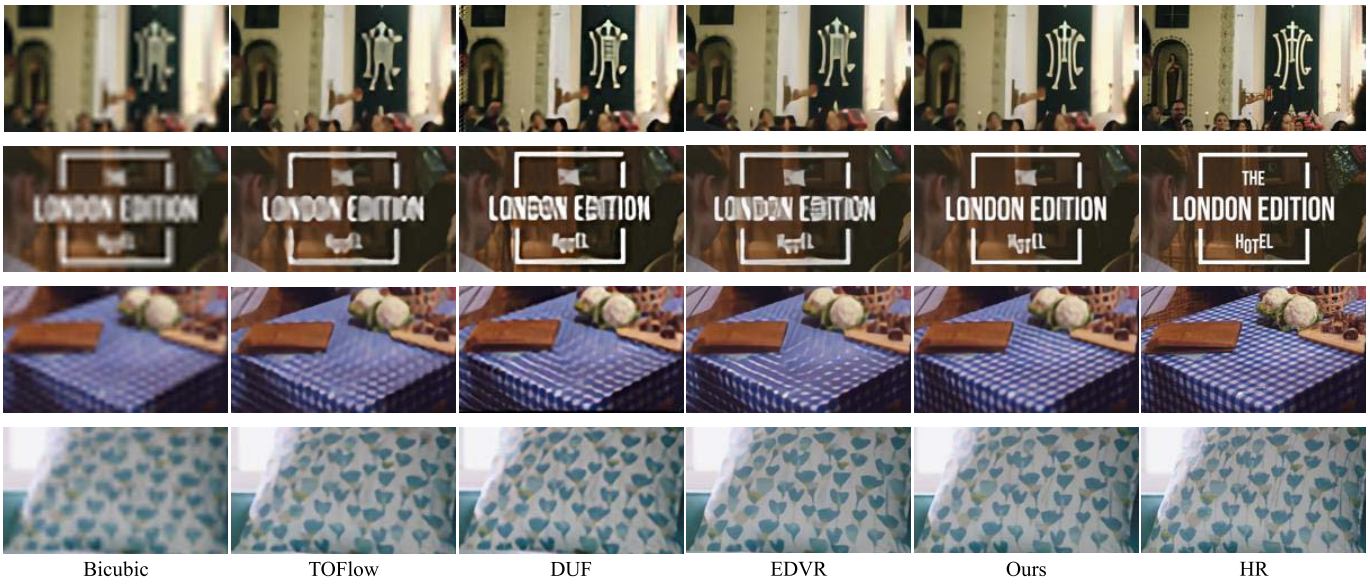


Fig. 7. Qualitative results on Vimeo-90K-T for 4× scaling factor.

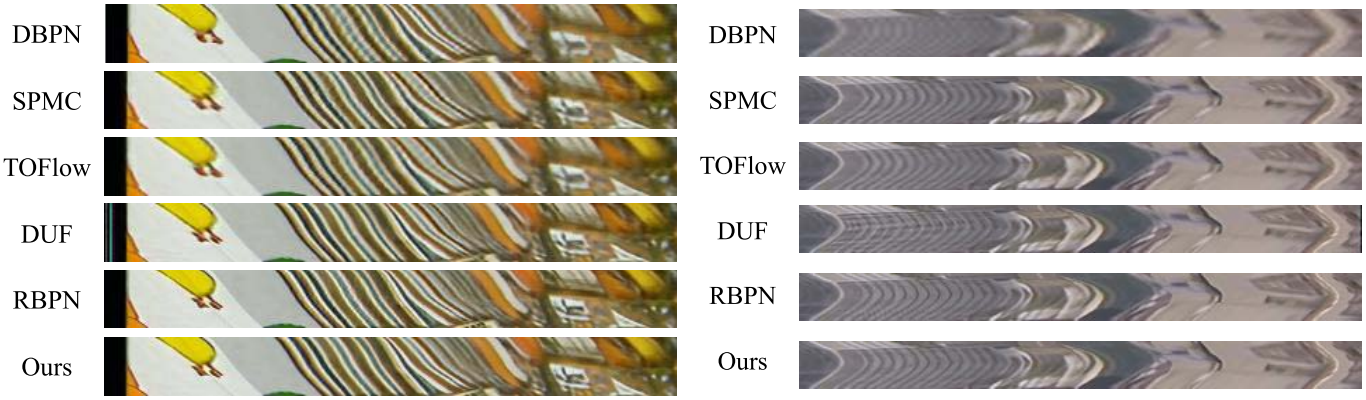


Fig. 8. Temporal consistency. (a) Temporal profiles for Calendar from Vid4. (b) Temporal profiles for City from Vid4(90 degree rotation for better view). Zoom in for best view.

It is obvious that our method is the unique approach to restore the abundant details and clean edges. The visual results on the Vimeo-90K-T dataset also demonstrates the superior visual quality achieved by our framework. The first row presented in Figure 7 shows that only our method can correctly and clearly recover the sign on the wall. In the second row, the text "THE LONDON EDITION HOTEL" reconstructed by our method has significantly sharper edges than previous methods. For the last two testing images, our results are also visually closest to the HR ground-truth frames.

3) Temporal Consistency Comparison: In addition to the quantitative and qualitative frame evaluations, we also test the inter-frame visual consistency of the generated SR videos. Temporal motion-based video integrity evaluation index (T-MOVIE) [46] is a widely-used way to evaluate the video temporal consistency. It can be observed in Figure 9 that our method outperforms most existing methods by a notable margin, which means that the results generated by our network are temporally more consistent. Similarly, temporal file [1] can also demonstrate our superior visual performance, which is assessed by taking the same horizontal or vertical row of pixels from a number of frames in the video and stacking them vertically or horizontally into a new image. From the exemplar results shown in Figure 8, we can see the results of SISR method DBPN have significant jitter and jagged lines, which indicates video content flickering across different frames. This is because SISR method separately estimates each output frame without considering temporal information. The VSR baseline methods SPMC, TOFlow, DUF and RBPN produce clearer results than DBPN, however they still have varying degree of flickering artifacts. Particularly, the blue line in the DUF result from Figure 8(a) demonstrates severe border effects. In contrast, our method produces the results with richest texture and edge information that faithfully describes the finer details in each image. Furthermore, VMAF[2] is a perceptual video quality assessment algorithm developed by Netflix [47]. From Table IV, we can observe

[2]It's worth noting that VMAF has an underlying assumption on the subject viewing distance and display size(1080p). The scores on low-resolution videos(720p) will look high.

TABLE IV

VMAF COMPARISON OF THE COMPARED METHODS. RESULTS ARE EVALUATED ON VID4. RED INDICATES THE BEST PERFORMANCE

Algorithm	Clip Name/Resolution			
	Calendar	City	Foliage	Walk
	720 × 576	704 × 576	720 × 480	720 × 480
SPMC	96.77	98.17	97.71	98.32
TOFlow	95.95	97.71	95.56	96.76
DUF	98.43	98.55	98.60	99.53
EDVR	98.68	98.63	98.40	99.67
RBPN	98.67	98.71	98.43	99.56
Ours	99.02	98.73	98.52	99.60

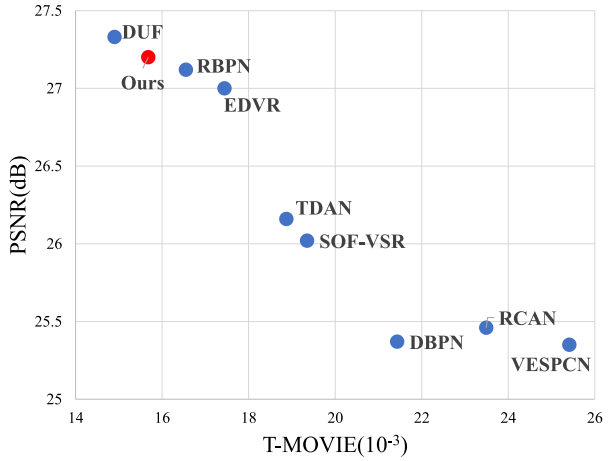


Fig. 9. Performance v.s. T-MOVIE of the compared methods. Results are evaluated on Vid4.

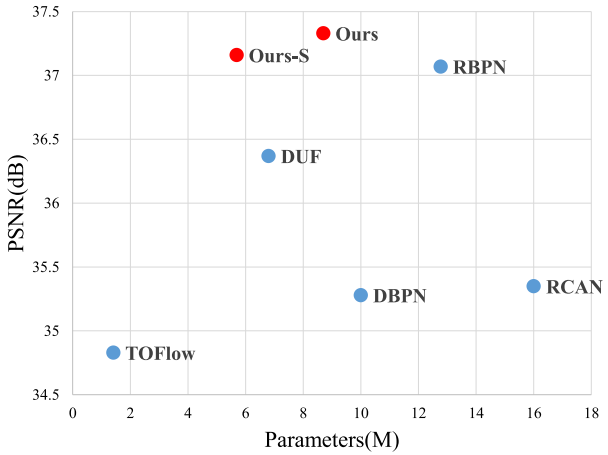


Fig. 10. Performance v.s. parameter number of the compared methods. Results are evaluated on Vimeo-90K-T.

TABLE V

COMPARISON OF DIFFERENT ALIGNMENT MODULES ON VID4

Alignment	-	OpFlow [25]	TDA [16]	PCD [17]	MDD(Ours)
Parameters(M)	4.21	5.43	4.90	5.73	5.69
PSNR(dB)	26.86	26.93	26.98	27.01	27.15

TABLE VI

COMPARISON OF DIFFERENT FUSION STRATEGIES ON VID4

Fusion strategy	DF [25]	3DF [1]	SFT (Ours)
PSNR(dB)	26.91	26.97	27.15

that our model provides a more consistent quality with human visual system in practical applications compared to other state-of-the-art methods. More video results can be found at https://github.com/XuWWJ/Video_Exapmles.

4) Model Efficiency Comparison: The model efficiency (the number of parameters, FLOPs, and runtime) are evaluated in Table I. As compared with all other SR methods, our model achieves a good trade-off between computational cost and performance. Specifically, Figure 10 shows the comparison results of the parameter number vs. reconstruction performance between the proposed method and other baseline methods. We use the PSNR results of Vimeo-90K-T dataset to demonstrate the performance of each model. DBPN and RCAN are the two best-performing SISR methods. They both have large model sizes, with parameter number up to more than 10 million. However, they have inferior performance for VSR as they ignore the video temporal information. The VSR methods in general achieve superior performance with much more lightweight models than SISR methods. RBPN achieves the highest PSNR value among all baseline VSR models with much more parameters than TOFlow and DUF. Compared to RBPN, our two models, Ours-S and Ours achieve much higher PSNR scores of 37.15 dB and 37.33 dB respectively, with about 50% fewer parameters. The parameter number of DUF is comparable to our two models but its PSNR value is about 0.8 dB lower than our two models.

D. Ablation Studies

To further investigate the proposed model, we conduct ablative experiments such as replacing or removing the key components of our framework. In this section, we present ablation results using our small model (Ours-S) on the Vid4 dataset to justify our design choices.

a) MDD Alignment Module: To verify the effectiveness of our MDD, we replace it by other excellent temporal alignment modules including optical flow based alignment module like OpFlow [25] and deformable convolution based module such as TDA [16] and PCD [17]. The last row of Table V lists the results of all testings. We can observe that without alignment, our model achieves the lowest PSNR score of 26.86 dB among the testings. Equipped with OpFlow, the performance of the model (OpFlow) boosts to 26.93 dB with a gain of 0.07 dB, verifying the effectiveness of temporal feature alignment for VSR. Compared to the OpFlow that has inaccurate optical flow estimation when suffering from large and complicate motions, the DConv-based TDA and PCD achieve much better performance with a gain of 0.05 dB and 0.08 dB, respectively. Our MDD module achieves the best PSNR score of 27.15 dB, significantly outperforming the second-best performing PCD by 0.14 dB. This is because the MDRB module in MDD can effectively extract multi-scale contextual features without blurring object boundary details compared to PCD. Furthermore, we make a comparison about parameter numbers, where the second row demonstrates that our MDD also has less parameters than PCD.

b) Multi-Stage Fusion Strategy: To validate the effectiveness of our multi-stage fusion module, we remove the SFT layers in f_{MFFN} and leverage the representative one-stage fusion mechanism direct fusion (DF) [25] and 3d convolution fusion (3DF) [1] for comparison.

TABLE VII
EFFECTIVENESS OF KEY COMPONENTS IN THE FRAMEWORK ON VID4

Baseline	Attention	Feedback	PSNR (dB)
✓			27.04
✓	✓		27.09
✓		✓	27.11
✓	✓	✓	27.15

TABLE VIII
EFFECTIVENESS OF DIFFERENT INPUT FRAME NUMBER N_f ON VID4

$N_f = 2N + 1$	PSNR(dB)/SSIM
1	25.39/0.7316
3	26.76/0.8030
5	27.02/0.8140
7	27.15/0.8197

DF [25] concatenates the aligned features of multiple frames and then conducts one-stage fusion via 2d convolution. 3DF [1] directly leverages 3d convolutions to extract spatio-temporal features and conduct one-stage fusion. As demonstrated by the Table VI, our multi-stage fusion strategy achieves the best PSNR score of 27.15 dB with a significant gain of 0.18 dB compared to the second-best performing 3DF, which demonstrates the effectiveness of our multi-stage fusion that can generate more accurate reconstruction results through progressively enhancing the features of the reference frame at each stage of the feature fusion process.

Effectiveness of Attention and Feedback. We further validate the necessity of the key components including attention module and feedback mechanism in our network. As listed by the top row of Table VII, the baseline model that removes our attention module and feedback mechanism achieves a PSNR score of 27.04 dB, which is much lower compared to our final model with a score of 27.15 dB. Even so, our baseline model still outperforms several early VSR methods such as TOFlow (25.89 db) and FRVSR (26.69 dB). Then, we add the attention module into the base model and the PSNR score is 27.09 dB, which has a gain of 0.05 dB than the baseline model. This verifies that refining the aligned features by the attention mechanism can further boost the performance. Next, we introduce the feedback mechanism into the base model, yielding a PSNR score of 27.11 dB with a gain of 0.07 dB against the baseline, which validates the effectiveness of the feedback strategy that leverages the high-level features to enhance the low-level features for better reconstruction accuracy. Moreover, with both attention module and feedback mechanism, our final model performs the best with a PSNR score of 27.15 dB, outperforming the baseline by 0.11 dB.

c) Effectiveness of Input Frame Number: In this subsection, we further explore the influence of input frames by leveraging different number of frames to train our network. It can be observed from Table VIII that with increase of the input frame number, the reconstruction performance is significantly improved. When taking 1 frame as input, the network does not utilize temporal information and is degenerated to an SISR model. Thus, the PSNR/SSIM values decrease extremely

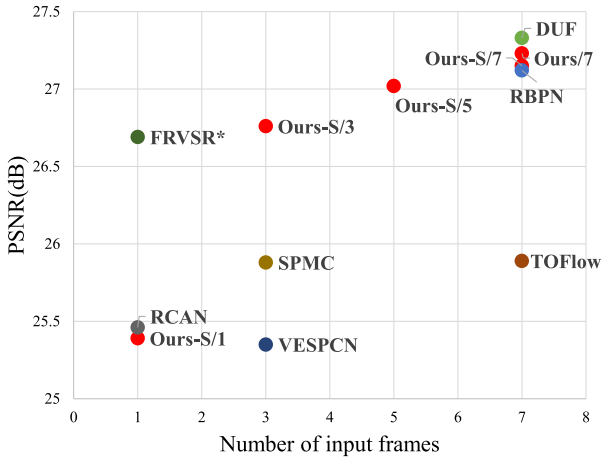


Fig. 11. Comparison results of different SR methods with a different number of input frames. 'Ours/N_f' denotes our methods with different number of frames $N_f = 1, 3, 5, 7$. Results are evaluated on Vid4. Note that FRVSR* is a temporal recurrent method.

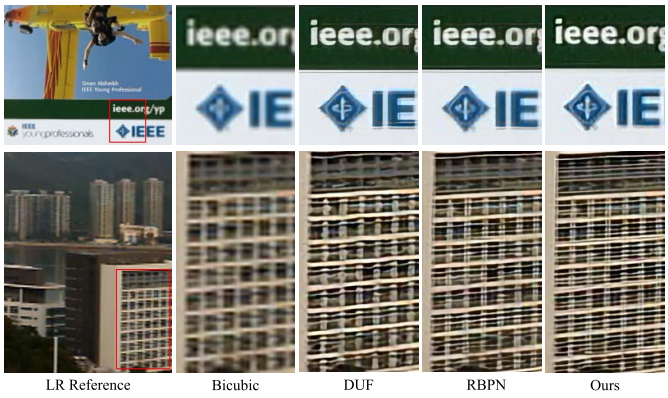


Fig. 12. Visual results on real-world video examples from *ieee* and *bldg* sequences.

compared to others. As the input frame number continues to increase from 3 to 7, the reconstruction accuracy is becoming better and better. This phenomenon is consistent with common sense, because more well-aligned supporting frames contain more rich supplementary information that is very helpful to restore the structure details of the reference frame. From Figure 11, we can observe clearly that our method has a significant improvement when increasing from 1 to 7 frames. And the performance of Ours-s/3 is much better than TOFlow which uses 7 frames. In fact, with the increase of input frames, the reconstruction performance of all methods improves a lot since more valuable temporal information are available.

E. Real-World Examples

The LR frames in the above discussions are all synthesized through bicubic downsampling operation, which may not fully reflect the complicate and challenging real-world case. To further validate the robustness of our proposed method, we also conduct experiments on real-world video sequences captured using a hand-held cellphone camera from [48]. The HR video frames are not available and the degradation methods are unknown. We compare our method with DUF and RBPN on the bicubic degradation configuration. Figure 12 shows the SR results for the two sequences *ieee* and *bldg*. Our method

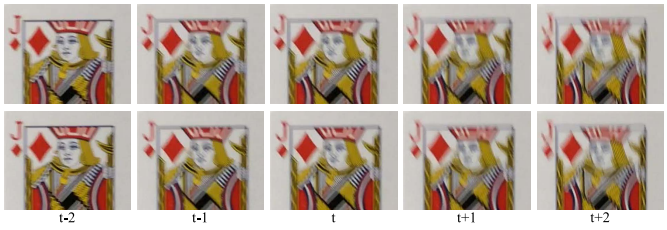


Fig. 13. Failure examples. Top is the LR reference and bottom is our SR results.

favorably recovers image characters (see top *ieee* results) and produces sharper edges (see bottom *bldg* results) than the compared state-of-the-art VSR networks. The comparison results show that our method have great potential to robustly handle even unknown degradation, which further demonstrates the superiority of the proposed framework over the state-of-the-arts for VSR.

F. Failure Cases

The LR video frames in real-world scenarios are not only limited to low resolution but also suffer from motion blur or low-frame rate due to shutter speed and exposure time of camera sensors. However, it is impossible to consider all these challenging factors for VSR modeling, which makes it very challenging to accurately reconstruct the structure details of the LR frames. One failure case of our method is shown in Figure 13. We can see that the LR reference suffers from severe motion blur and our method fails to recover the clear face in the poker card since we don't take the deblur step into consideration. Therefore, to be better adapted to real-world VSR, it is worth to address the joint video enhancement issue such as spatial and temporal SR or deblur and spatial SR [12].

V. Conclusion

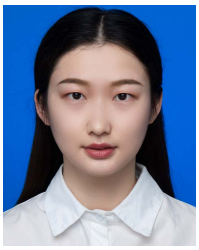
In this paper, we propose a novel VSR method called Multi-Stage Feature Fusion Network. We align the features of the reference frame and the supporting frames with multi-scale dilated deformable convolutions, and use the aligned feature as an conditional input to be fused with the feature of the LR reference frame at different stages of a network backbone. By doing this, we can relieve the discrepancy between the feature of the LR reference frame the fused features of a video sequence. Extensive experiments demonstrate that our method achieves state-of-the-art performance on several benchmark VSR datasets. For future work, we plan to dynamically select the optimal range of the supporting frames to be fused with the reference frame, and achieve temporally dynamic multi-stage feature fusion.

References

[1] J. Caballero *et al.*, "Real-time video super-resolution with spatio-temporal networks and motion compensation," in *Proc. CVPR*, 2017, pp. 4778–4787.
[2] X. Tao, H. Gao, R. Liao, J. Wang, and J. Jia, "Detail-revealing deep video super-resolution," in *Proc. ICCV*, 2017, pp. 4472–4480.
[3] M. Haris, G. Shakhnarovich, and N. Ukita, "Recurrent back-projection network for video super-resolution," in *Proc. CVPR*, 2019, pp. 3897–3906.
[4] C. Liu and D. Sun, "A Bayesian approach to adaptive video super resolution," in *Proc. CVPR*, 2011, pp. 209–216.
[5] P. Yi, Z. Wang, K. Jiang, J. Jiang, and J. Ma, "Progressive fusion video super-resolution network via exploiting non-local spatio-temporal correlations," in *Proc. ICCV*, 2019, pp. 3106–3115.
[6] W. Shi *et al.*, "Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 1874–1883.
[7] J. Kim, J. K. Lee, and K. M. Lee, "Deeply-recursive convolutional network for image super-resolution," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 1637–1645.
[8] C. Ledig *et al.*, "Photo-realistic single image super-resolution using a generative adversarial network," in *Proc. CVPR*, 2017, pp. 4681–4690.
[9] B. Lim, S. Son, H. Kim, S. Nah, and K. Mu Lee, "Enhanced deep residual networks for single image super-resolution," in *Proc. CVPRW*, 2017, pp. 136–144.
[10] Z. Li, J. Yang, Z. Liu, X. Yang, G. Jeon, and W. Wu, "Feedback network for image super-resolution," in *Proc. CVPR*, 2019, pp. 3867–3876.
[11] K. Zhang, W. Zuo, and L. Zhang, "Deep plug-and-play super-resolution for arbitrary blur kernels," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2019, pp. 1671–1681.
[12] S. Nah *et al.*, "Ntire 2019 challenge on video super-resolution: Methods and results," in *Proc. CVPRW*, 2019, pp. 1985–1995.
[13] D. Ren, K. Zhang, Q. Wang, Q. Hu, and W. Zuo, "Neural blind deconvolution using deep priors," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2020, pp. 3338–3347.
[14] D. Ren, W. Zuo, D. Zhang, L. Zhang, and M.-H. Yang, "Simultaneous fidelity and regularization learning for image restoration," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 1, pp. 284–299, Jan. 2019.
[15] 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. ECCV*, 2018, pp. 286–301.
[16] Y. Tian, Y. Zhang, Y. Fu, and C. Xu, "TDAN: Temporally deformable alignment network for video super-resolution," in *Proc. CVPR*, 2020, pp. 3360–3369.
[17] X. Wang, K. C. Chan, K. Yu, C. Dong, and C. C. Loy, "EDVR: Video restoration with enhanced deformable convolutional networks," in *Proc. CVPRW*, 2019, pp. 1954–1963.
[18] Y. Huang, W. Wang, and L. Wang, "Bidirectional recurrent convolutional networks for multi-frame super-resolution," in *Proc. NIPS*, 2015, pp. 235–243.
[19] D. Liu *et al.*, "Robust video super-resolution with learned temporal dynamics," in *Proc. ICCV*, 2017, pp. 2507–2515.
[20] Y. Jo, S. Wug Oh, J. Kang, and S. J. Kim, "Deep video super-resolution network using dynamic upsampling filters without explicit motion compensation," in *Proc. CVPR*, 2018, pp. 3224–3232.
[21] X. Zhu, H. Hu, S. Lin, and J. Dai, "Deformable ConVnets v2: More deformable, better results," in *Proc. CVPR*, 2019, pp. 9308–9316.
[22] M. S. Sajjadi, R. Vemulapalli, and M. Brown, "Frame-recurrent video super-resolution," in *Proc. CVPR*, 2018, pp. 6626–6634.
[23] S. Xingjian, Z. Chen, H. Wang, D.-Y. Yeung, W.-K. Wong, and W.-C. Woo, "Convolutional LSTM network: A machine learning approach for precipitation nowcasting," in *Proc. NIPS*, 2015, pp. 802–810.
[24] A. Kappeler, S. Yoo, Q. Dai, and A. K. Katsaggelos, "Video super-resolution with convolutional neural networks," *IEEE Trans. Comput. Imag.*, vol. 2, no. 2, pp. 109–122, Jun. 2016.
[25] L. Wang, Y. Guo, Z. Lin, X. Deng, and W. An, "Learning for video super-resolution through hr optical flow estimation," in *Proc. ACCV*. Perth, WA, Australia: Springer, 2018, pp. 514–529.
[26] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," 2015, *arXiv:1502.03167*. [Online]. Available: http://arxiv.org/abs/1502.03167
[27] V. Dumoulin, J. Shlens, and M. Kudlur, "A learned representation for artistic style," 2016, *arXiv:1610.07629*. [Online]. Available: http://arxiv.org/abs/1610.07629
[28] G. Ghiasi, H. Lee, M. Kudlur, V. Dumoulin, and J. Shlens, "Exploring the structure of a real-time, arbitrary neural artistic stylization network," 2017, *arXiv:1705.06830*. [Online]. Available: http://arxiv.org/abs/1705.06830
[29] X. Huang and S. Belongie, "Arbitrary style transfer in real-time with adaptive instance normalization," in *Proc. ICCV*, 2017, pp. 1501–1510.
[30] H. De Vries, F. Strub, J. Mary, H. Larochelle, O. Pietquin, and A. C. Courville, "Modulating early visual processing by language," in *Proc. NIPS*, 2017, pp. 6594–6604.
[31] E. Perez, H. de Vries, F. Strub, V. Dumoulin, and A. Courville, "Learning visual reasoning without strong priors," 2017, *arXiv:1707.03017*. [Online]. Available: http://arxiv.org/abs/1707.03017

[32] E. Perez, F. Strub, H. De Vries, V. Dumoulin, and A. Courville, "Film: Visual reasoning with a general conditioning layer," in *Proc. AAAI*, 2018, pp. 3942–3951.
[33] X. Wang, K. Yu, C. Dong, and C. Change Loy, "Recovering realistic texture in image super-resolution by deep spatial feature transform," in *Proc. CVPR*, 2018, pp. 606–615.
[34] J. Dai *et al.*, "Deformable convolutional networks," in *Proc. ICCV*, 2017, pp. 764–773.
[35] G. Bertasius, L. Torresani, and J. Shi, "Object detection in video with spatiotemporal sampling networks," in *Proc. ECCV*, 2018, pp. 331–346.
[36] Y. Zhao, Y. Xiong, and D. Lin, "Trajectory convolution for action recognition," in *Proc. NIPS*, 2018, pp. 2204–2215.
[37] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *Proc. ECCV*, 2018, pp. 801–818.
[38] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. CVPR*, 2016, pp. 770–778.
[39] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam, "Rethinking atrous convolution for semantic image segmentation," 2017, *arXiv:1706.05587*. [Online]. Available: http://arxiv.org/abs/1706.05587
[40] M. Haris, G. Shakhnarovich, and N. Ukita, "Deep back-projection networks for super-resolution," in *Proc. CVPR*, 2018, pp. 1664–1673.
[41] T. Xue, B. Chen, J. Wu, D. Wei, and W. T. Freeman, "Video enhancement with task-oriented flow," *Int. J. Comput. Vis.*, vol. 127, no. 8, pp. 1106–1125, 2019.
[42] 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.
[43] W.-S. Lai, J.-B. Huang, N. Ahuja, and M.-H. Yang, "Fast and accurate image super-resolution with deep Laplacian pyramid networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 11, pp. 2599–2613, Nov. 2019.
[44] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on imagenet classification," in *Proc. ICCV*, 2015, pp. 1026–1034.
[45] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*. [Online]. Available: http://arxiv.org/abs/1412.6980
[46] K. Seshadrinathan and A. C. Bovik, "Motion tuned spatio-temporal quality assessment of natural videos," *IEEE Trans. Image Process.*, vol. 19, no. 2, pp. 335–350, Feb. 2010.
[47] Z. Li, A. Aaron, I. Katsavounidis, A. Moorthy, and M. Manohara, "Toward a practical perceptual video quality metric," *Netflix Tech Blog*, vol. 6, no. 2, 2016.
[48] R. Liao, X. Tao, R. Li, Z. Ma, and J. Jia, "Video super-resolution via deep draft-ensemble learning," in *Proc. ICCV*, 2015, pp. 531–539.

Huihui Song received the B.S. degree in technology and science of electronic information from the Ocean University of China in 2008, the master's degree in communication and information system from the University of Science and Technology of China (USTC) in 2011, and the Ph.D. degree in geography and resource management from The Chinese University of Hong Kong in 2014. She is currently a Professor with the Jiangsu Key Laboratory of Big Data Analysis Technology, Nanjing University of Information Science and Technology, Nanjing, China. Her research interests include remote sensing image processing and image fusion.



Wenjie Xu is currently pursuing the M.S. degree with the School of Automation, Nanjing University of Information Science and Technology, Nanjing, China. Her current research interest includes image/video super-resolution algorithms.

Dong Liu received the Ph.D. degree in computer science from the Harbin Institute of Technology. He was with RetailNext Inc., focusing on designing the lightweight single-shot multi-task neural network architectures for real time retail analytics on embedded devices. Prior to that, he was a Postdoctoral Research Scientist with the DVMM Laboratory, Columbia University, New York, NY, USA. Over the years, he has developed innovative solutions for image/video content analysis, multimodal information fusion, large-scale visual search/recommendation, and fundamental machine learning methods. He is currently a Senior Research Scientist with Netflix Research. His research interests include computer vision, machine learning, and multimedia information retrieval.

Bo Liu received the Ph.D. degree from the Computer Science Department, Rutgers, The State University of New Jersey, in 2018. He worked as a Research Staff with The Hong Kong Polytechnic University. He is currently a Research Scientist with JD Finance America Corporation. His other previous employments include Siemens Healthineers, GE Global Research, and Microsoft Research Asia. His current research interests include machine learning, computer vision, and data analytics.

Qingshan Liu (Senior Member, IEEE) received the M.S. degree from the Department of Auto Control, Southeast University, Nanjing, in 2000, and the Ph.D. degree from the National Laboratory of Pattern Recognition, Chinese Academic of Science, Beijing, China, in 2003. He was an Assistant Research Professor with the Computational Biomedicine Imaging and Modeling Center, Department of Computer Science, Rutgers, The State University of New Jersey, New Brunswick, NJ, USA, from 2010 to 2011. Before he joined Rutgers University, he was an Associate Professor with the National Laboratory of Pattern Recognition, Chinese Academic of Science, and an Associate Researcher with the Multimedia Laboratory, The Chinese University of Hong Kong, Hong Kong, from 2004 and 2005. He is currently a Professor with the School of Automation, Nanjing University of Information Science and Technology, Nanjing, China. His current research interests are image and vision analysis, including face image analysis, graph and hypergraph-based image and video understanding, medical image analysis, and event-based video analysis. He was a recipient of the President Scholarship of the Chinese Academy of Sciences in 2003.

Dimitris N. Metaxas (Fellow, IEEE) received the B.E. degree from the National Technical University of Athens, Greece, in 1986, the M.S. degree from The University of Maryland, in 1988, and the Ph.D. degree from the University of Toronto, in 1992. He is currently a Professor with the Computer Science Department, Rutgers University. He is also directing the Computational Biomedicine Imaging and Modeling Center (CBIM). He has been conducting research toward the development of formal methods upon which computer vision, computer graphics, and medical imaging can advance synergistically.