Artificial Intelligence Review
https://doi.org/10.1007/s10462-022-10147-y

Video super-resolution based on deep learning: a comprehensive survey

Hongying Liu[1,2] · Zhubo Ruan[1] · Peng Zhao[1] · Chao Dong[3] · Fanhua Shang[1,2] · Yuanyuan Liu[1] · Linlin Yang[1] · Radu Timofte[4,5]

Abstract
Video super-resolution (VSR) is reconstructing high-resolution videos from low resolution ones. Recently, the VSR methods based on deep neural networks have made great progress. However, there is rarely systematical review on these methods. In this survey, we comprehensively investigate 37 state-of-the-art VSR methods based on deep learning. It is well known that the leverage of information contained in video frames is important for video super-resolution. Thus we propose a taxonomy and classify the methods into seven subcategories according to the ways of utilizing inter-frame information. Moreover, descriptions on the architecture design and implementation details are also included. Finally, we summarize and compare the performance of the representative VSR methods on some benchmark datasets. We also discuss the applications, and some challenges, which need to be further addressed by researchers in the community of VSR. To the best of our knowledge, this work is the first systematic review on VSR tasks, and it is expected to make a contribution to the development of recent studies in this area and potentially deepen our understanding of the VSR techniques based on deep learning.

Keywords Video super-resolution · Deep learning · Convolutional neural networks · Inter-frame information

✉ Fanhua Shang
fhshang@xidian.edu.cn

[1] Key Laboratory of Intelligent Perception and Image Understanding of Ministry of Education, School of Artificial Intelligence, Xidian University, Xi'an, China

[2] Peng Cheng Laboratory, Shenzhen, China

[3] The Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, Shenzhen, China

[4] ETH Zurich, Zurich, Switzerland

[5] University of Wurzburg, Würzburg, Germany

Published online: 01 April 2022

1 Introduction

Super-resolution (SR) aims at recovering a high-resolution (HR) image or multiple images from the corresponding low-resolution (LR) counterparts. It is a classic and challenging problem in computer vision and image processing, and it has extensive real-world applications, such as medical imagery reconstruction (Peng et al. 2020), remote sensing (Luo et al. 2017), and panorama video super-resolution (Fakour-Sevom et al. 2018; Liu et al. 2020b), surveillance systems (Deshmukh and Rani 2019), and high-definition television (Patti et al. 1997). With the advent of the 5th generation mobile communication technology, large-sized images or videos can be transmitted within a shorter time. Meanwhile, with the popularity of high-definition (HD) and ultra-high-definition (UHD) display devices, video super-resolution is attracting more attention.

Video is one of the most common multimedia in our daily life, and thus super-resolution of low-resolution videos has become very important. In general, image super-resolution methods process a single image at a time, while video super-resolution algorithms deal with multiple successive images/frames at a time so as to utilize relationship within frames to super-resolve the target frame. In a broad sense, video super-resolution (VSR) can be regarded as an extension of image super-resolution and can be processed by image super-resolution algorithms frame by frame. However, the SR performance is not always satisfactory as artifacts and jams may be brought in, which causes unwanted temporal incoherence within frames.

In recent years, many video super-resolution algorithms have been proposed. They mainly fall into two categories: traditional methods and deep learning based methods. For some traditional methods, the motions are simply estimated by affine models as in Schultz and Stevenson (1996). In Protter et al. (2009) and Takeda et al. (2009), they adopt non-local mean and 3D steering kernel regression for video super-resolution, respectively. Liu and Sun (2014) proposed a Bayesian approach to simultaneously estimate underlying motion, blur kernel, and noise level for reconstructing high-resolution frames. In Ma et al. (2015), the expectation maximization (EM) method is adopted to estimate the blur kernel, and guide the reconstruction of high-resolution frames. However, these explicit models of high-resolution videos are still inadequate to fit various scenes in videos.

With the great success of deep learning in a variety of areas (Zhang et al. 2021), super-resolution algorithms based on deep learning are studied extensively. Many video super-resolution methods based on deep neural networks such as convolutional neural network (CNN), generative adversarial network (GAN) and recurrent neural network (RNN) have been proposed. Generally, they employ a large number of both LR and HR video sequences to input the neural network for inter-frame alignment, feature extraction/fusion, and then to produce the high-resolution sequences for the corresponding low-resolution video sequences. The pipeline of most video super-resolution methods mainly includes one alignment module, one feature extraction and fusion module, and one reconstruction module, as shown in Fig. 1. Because of the nonlinear learning capability of deep neural networks, the deep learning based methods usually achieve good performance on many public benchmark datasets.

So far, there are few works about the overview on video super-resolution tasks, though many works (Wang et al. 2021a; Singh and Singh 2020; Yang et al. 2019) on the investigation of single image super-resolution have been published. Daithankar and Ruikar (2020) presented a brief review on many frequency-spatial domain methods, while the deep learning methods are rarely mentioned. Unlike the previous work, we provide a comprehensive

investigation on deep learning techniques for video super-resolution in recent years. It is well known that the main difference between video super-resolution and image super-resolution lies in the processing of inter-frame information. How to effectively leverage the information from neighboring frames is critical for VSR tasks. We focus on the ways of utilizing inter-frame information for various deep learning based methods.

The contributions of this work are mainly summarized as follows. (1) We review recent works and progresses on developing techniques for deep learning based video super-resolution. To the best of our knowledge, this is the first comprehensive survey on deep learning based VSR methods. (2) We propose a taxonomy for deep learning based video super-resolution methods by categorizing their ways of utilizing inter-frame information and illustrate how the taxonomy can be used to categorize existing methods. (3) We summarize the performance of state-of-the-art methods on some public benchmark datasets, and list the applications of VSR algorithms in various areas. (4) We further discuss some challenges and perspectives for video super-resolution tasks.

The rest of the paper is organized as follows. In Sect. 2, we briefly introduce the background of video super-resolution. Section 3 shows our taxonomy for recent works. In Sects. 4 and 5, we describe the video super-resolution methods with and without alignment, respectively, according to the taxonomy. In Sect. 6, the performance of state-of-the-art methods is analyzed quantitatively. In Sect. 7, we discuss the challenges and prospective trends in video super-resolution. Finally, we conclude this work in Sect. 8.

2 Background

Video super-resolution stems from image super-resolution, and it aims at restoring high-resolution videos from multiple low-resolution frames. However, the difference between video and image super-resolution techniques is also obvious, that is, the former usually takes advantage of inter-frame information. Besides the RGB color space, the YUV including YCbCr color space is also widely used for VSR. $I_i \in \mathbb{R}^{H\times W\times 3}$ denotes the i-th frame in an LR video sequence I, and $\hat{I}_i \in \mathbb{R}^{sH\times sW\times 3}$ is the corresponding HR frame, where s is the scale factor, e.g., $s = 2$, 4 or 8. And $\{\hat{I}_j\}_{j=i-N}^{i+N}$ is a set of $2N + 1$ HR frames for the center frame $\hat{I}_i$, where N is the temporal radius. Then the degradation process of HR video sequences can be formulated as follows:

$$I_i = \phi(\hat{I}_i, \{\hat{I}_j\}_{j=i-N}^{i+N}; \theta_\alpha) \tag{1}$$

where $\phi(\cdot;\cdot)$ is the degradation function, and the parameter θ_α represents various degradation factors such as noise, motion blur and downsampling factors. In most existing works (Liu and Sun 2014; Ma et al. 2015; Farsiu et al. 2004; Pan et al. 2020), the degradation process is expressed as:

$$I_j = DBE_{i\to j}\hat{I}_i + n_j \tag{2}$$

where D and B are the down-sampling and blur operations, n_j denotes image noise, and $E_{i\to j}$ is the warping operation based on the motion from $\hat{I}_i$ to $\hat{I}_j$.

In practice, it is easy to obtain LR image I_j, but the degradation factors, which may be quite complex or probably a combination of several factors, are unknown. Different from single image super-resolution (SISR) aiming at solving a single degraded image, VSR needs to deal

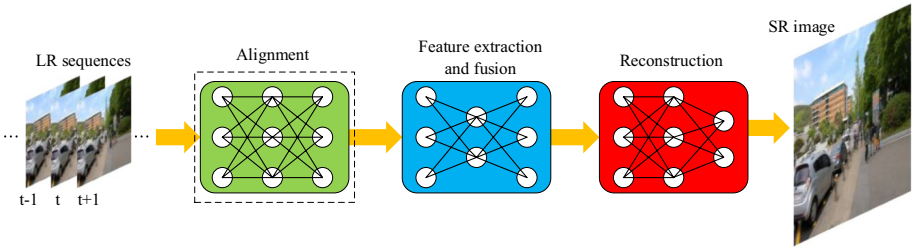

Fig. 1 The general pipeline of deep learning methods for VSR tasks. Note that the inter-frame alignment module can be either traditional methods or deep CNNs, while both the feature extraction & fusion module and the upsampling module usually utilize deep CNNs. The dashed line box means that the module is optional

with degraded video sequences, and recovers the corresponding HR video sequences, which should be as close to the ground truth (GT) videos as possible. Specifically, a VSR algorithm may use similar techniques to SISR for processing a single frame (spatial information), while it has to take relationships among frames (temporal information) into consideration to ensure motion consistency of the video. The super-resolution process, namely the reverse process of Eq. (1), can be formulated as follows:

$$\tilde{I}_i = \phi^{-1}(I_i, \{I_j\}_{j=i-N}^{i+N}; \theta_\beta) \tag{3}$$

where $\tilde{I}_i$ denotes the estimation of the GT (i.e., $\hat{I}_i$), and θ_β is the model parameter.

Like SISR, video quality is mainly evaluated by calculating peak signal-noise ratio (PSNR) and structural similarity index (SSIM). These indexes measure the difference of pixels and similarity of structures between two images, respectively. PSNR of one SR frame is defined as:

$$PSNR = 10 \log_{10} \left(\frac{L^2}{MSE} \right) \tag{4}$$

where L represents the maximum range of color value, which is usually 255, and the mean squared error (MSE) is defined as:

$$MSE = \frac{1}{N} \sum_{i=1}^{N} (\hat{I}_i - \tilde{I}_i)^2 \tag{5}$$

where N denotes the total number of pixels in an image or a frame, $\hat{I}$ and $\tilde{I}$ are the ground truth HR frame and the SR recovered frame, respectively. A higher value of PSNR generally means superior quality. In addition, SSIM is defined as:

$$SSIM(\hat{I}, \tilde{I}) = \frac{2u_{\hat{I}}u_{\tilde{I}} + k_1}{u_{\hat{I}}^2 + u_{\tilde{I}}^2 + k_1} \cdot \frac{2\sigma_{\hat{I}\tilde{I}} + k_2}{\sigma_{\hat{I}}^2 + \sigma_{\tilde{I}}^2 + k_2} \tag{6}$$

where $u_{\hat{I}}$ and $u_{\tilde{I}}$ represent the mean values of the images $\hat{I}$ and $\tilde{I}$, respectively. k_1 and k_2 are constants, which are used to stabilize the calculation and are usually set to 0.01 and 0.03, respectively. $\sigma_{\hat{I}}$ and $\sigma_{\tilde{I}}$ denote the standard deviations, and $\sigma_{\hat{I}\tilde{I}}$ denotes the covariance.

3 Video super-resolution methods

As the videos are a recording of moving visual images and sound, the methods for video super-resolution learn from existing single image super-resolution methods. There are many deep learning based image super-resolution methods such as Super-Resolution using deep Convolutional Neural Networks (SRCNN) (Dong et al. 2014), Fast Super-Resolution Convolutional Neural Networks (FSRCNN) (Dong et al. 2016), VDSR (Kim et al. 2016), Efficient Sub-Pixel Convolutional neural Network (ESPCN) (Shi et al. 2016), Residual Dense Network (RDN) (Zhang et al. 2018c), Residual Channel Attention Network (RCAN) (Zhang et al. 2018b), "Zero-Shot" Super-Resolution (ZSSR) (Shocher et al. 2018) and Super-Resolution using a Generative Adversarial Network (SRGAN) (Ledig et al. 2017). In 2016, based on SRCNN, Kappeler et al. (2016) presented a video super-resolution method with convolutional neural networks (VSRnet). So far, many video super-resolution algorithms have been proposed. In the following, we summarize the characteristics of the deep learning based methods for video super-resolution in recent years, as shown in Table 1.

Several studies such as Wang et al. (2019c), Jo et al. (2018) and Tian et al. (2020) on video super-resolution have indicated that the utilization of the information contained in frames greatly influences performance. The proper and adequate usage of such information can enhance the results of video super-resolution. Therefore, we build a taxonomy for existing video super-resolution methods according to their ways of the utilization of inter-frame information, as shown in Fig. 2.

As shown in Fig. 2 and Table 1, we categorize the existing methods into two main categorises: methods with alignment and methods without alignment, according to whether the video frames are explicitly aligned. We will present the methods in detail in the following sections.

Since all the methods are classified according to whether the frames are explicitly aligned and what techniques they are mainly used for alignment, other modules which they utilized for feature extraction, fusion, and reconstruction are ignored. These modules may be employed by multiple methods simultaneously. Therefore, some of the methods in our study are coupled. BasicVSR in the MEMC methods from the category of the methods with alignment adopts a typical bidirectional recurrent convolutional neural network (RCNN) as backbone. While the RCNN-based methods (e.g. BRCN, STCN, and RISTN), which are in the methods without alignment, mainly use RCNN to learn features. Similarly, VESR-Net in the DC category also uses a non-local block for feature learning as that of PFNL in the non-local category. Moreover, DSMC in the category of 3D convolution also utilizes a non-local block for global correlation computation. The category of 'other' includes the methods which adopt optical flow but without frame alignment, e.g., RBPN, and STARnet. Finally, the learned offsets by deformable convolution share similar patterns as those from the optical flow-based methods, and the deformable and flow-based alignments are strongly correlated. This was indicated in the work (Chan et al. 2021b).

Moreover, we have observed several trends in these recently proposed methods. (1) The diversification of methods. In the early years (2015–2017), most of the methods use frame alignment for VSR. Then since 2018, many different methods, especially which are the methods without alignment, have emerged, e.g., FFCVSR, DUF, RISTN, and PLNL. Some studies also indicate that both the methods with alignment and those without alignment can obtain sound performance. (2) The expansion of receptive field in methods. Earlier methods such as EDVR and RBPN mainly utilize certain numbers of input frames in

Table 1 Existing video super-resolution methods based on deep learning and their key strategies such as loss functions (see their source papers for the details of the loss functions)

Method	Year	Synonym	Type	Loss function	Align
Deep-DE (Liao et al. 2015)	ICCV 2015	Deep Draft-Ensemble Learning	MEMC	ℓ_1-norm loss with total variation regularization	✓
VSRnet (Kappeler et al. 2016)	TCI 2016	Video Super-Resolution with convolutional neural Networks		Mean Square Error (MSE) loss	✓
VESPCN (Caballero et al. 2017)	CVPR 2017	Video Efficient Sub-pixel Convolutional Network		MSE loss and Motion Compensation (MC) loss	✓
DRVSR (Tao et al. 2017)	ICCV 2017	Detail-Revealing deep Video Super-Resolution		MSE loss and MC loss	✓
RVSR (Liu et al. 2017)	ICCV 2017	Robust Video Super-Resolution		Spatial alignment loss and spatio-temporal adaptive loss	✓
FRVSR (Sajjadi et al. 2018)	CVPR 2018	Frame-Recurrent Video Super-Resolution		MSE loss and MC loss	✓
STTN (Kim et al. 2018a)	ECCV 2018	Spatio-Temporal Transformer Network		MSE loss and MC loss	✓
SOFVSR (Wang et al. 2019d)	ACCV 2018	Super-resolution Optical Flow for Video Super-Resolution		MSE loss and MC loss	✓
TOFlow (Xue et al. 2019)	IJCV 2019	video enhancement with Task-Oriented Flow		ℓ_1-norm loss	✓
MMCNN (Wang et al. 2019b)	TIP 2019	Multi-MemoryConvolutional Neural Network		MSE loss and MC loss	✓
MEMC-Net (Bao et al. 2021)	TPAMI 2019	Motion Estimation and Motion Compensation Network		Charbonnier (Cb) loss	✓
RRCN (Li et al. 2019a)	TIP 2019	Residual Recurrent Convolutional Network		MSE loss	✓
RTVSR (Bare et al. 2019)	Neurocomp. 2019	Real-Time Video Super-Resolution		MSE loss	✓
MultiBoot VSR (Kalarot and Porikli 2019)	CVPRW 2019	Multi-stage multi-reference Bootstrapping for Video Super-Resolution		Huber loss	✓
MuCAN (Li et al. 2020)	ECCV 2020	Multi-Correspondence Aggregation Network for Video Super-Resolution		Edge-aware loss	✓
TecoGAN (Chu et al. 2020)	ACMTOG 2020	Temporally coherent GAN		MSE loss and ping-pong loss etc.	✓
BasicVSR (Chan et al. 2021d)	CVPR 2021	search for essential components in Video Super-Resolution and beyond		Cb loss	✓

Table 1 (continued)

Method	Year	Synonym	Type	Loss function	Align
EDVR (Wang et al. 2019c)	CVPRW 2019	Enhanced Deformable convolutional networks for Video Restoration	DC	Cb loss	✓
DNLN (Wang et al. 2019a)	ACCESS 2019	Deformable Non-Local Network for Video Super-Resolution		ℓ_1-norm loss	✓
TDAN (Tian et al. 2020)	CVPR 2020	Temporally-Deformable Alignment Network for Video Super-Resolution		ℓ_1-norm loss	✓
D3Dnet (Ying et al. 2020)	SPL 2020	Deformable 3D Convolution for Video Super-Resolution		MSE loss	✓
VESR-Net (Chen et al. 2020)	ArXiv 2020	Video Enhancement and Super-Resolution Network		ℓ_1-norm loss	✓
VSRResFeatGAN (Lucas et al. 2019)	TIP 2019	Video Super-Resolution with Residual Networks	2D Conv	Adversarial loss; content loss; and perceptual loss	×
FFCVSR (Yan et al. 2019)	AAAI 2019	Frame and Feature-Context Video Super-Resolution		MSE loss	×
DUF (Jo et al. 2018)	CVPR 2018	video super-resolution network using Dynamic Upsampling Filters	3D Conv	Huber loss	×
FSTRN (Li et al. 2019b)	CVPR 2019	FastSpatio-Temporal Residual Network for Video Super-Resolution		Cb loss	×
3DSRnet (Kim et al. 2019)	ICIP 2019	3D Super-Resolution Network		MSE loss	×
DSMC (Liu et al. 2021b)	AAAI 2021	Dual Subnet and Multi-stage Communicated upsampling		Cb loss; perceptual loss; the dual loss	×

Table 1 (continued)

Method	Year	Synonym	Type	Loss function	Align
BRCN (Huang et al. 2015, 2018)	NIPS 2015/2018	video super-resolution via Bidirectional Recurrent Convolutional Networks	RCNN	MSE loss	×
STCN (Guo and Chao 2017)	AAAI 2017	Spatio-Temporal Convolutional Network for Video Super-Resolution		MSE loss	×
RISTN (Zhu et al. 2019b)	AAAI 2019	Residual Invertible Spatio-Temporal Network for Video Super-Resolution		MSE loss	×
RLSP (Fuoli et al. 2019a)	ICCVW 2019	video super-resolution through Recurrent Latent Space Propagation		MSE loss	×
RSDN (Isobe et al. 2020)	ECCV 2020	video super-resolution with Recurrent Structure-Detail Network		Cb loss	×
PFNL (Yi et al. 2019)	ICCV 2019	Progressive Fusion network via exploiting Non-Local spatio-temporal correlations	Non-Local	Cb loss	×
RBPN (Haris et al. 2019)	CVPR 2019	Recurrent Back-Projection Network	Other	ℓ_1-norm loss	×
STARnet (Haris et al. 2020)	CVPR 2020	Space-Time-Aware multi-Resolution network		Three losses	×
DNSTNet (Sun et al. 2020)	Neurocomp. 2020	video super-resolution via Dense Non-local Spatial-Temporal convolutional Network		ℓ_1-norm loss	×

Here, *MEMC* denotes motion estimation and motion compensation, *DC* is deformable convolution, 3*D Conv* is 3D convolution, and *RCNN* denotes recurrent convolutional neural networks

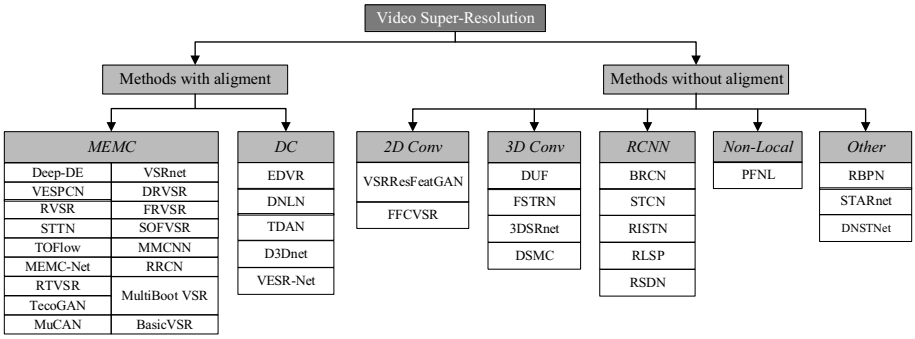

Fig. 2 A taxonomy for existing state-of-the-art video super-resolution methods. Here, *MEMC* stands for motion estimation and compensation methods, *DC* is deformable convolution methods, *3D Conv* is 3D convolution methods, and *RCNN* denotes recurrent convolutional neural network based methods. The links to these methods are as follows. MEMC: Deep-DE (Liao et al. 2015),VSRnet (Kappeler et al. 2016), VESPCN (Caballero et al. 2017), DRVSR (Tao et al. 2017), RVSR (Liu et al. 2017), FRVSR (Sajjadi et al. 2018), STTN (Kim et al. 2018a), SOFVSR (Wang et al. 2019d), TOFlow (Xue et al. 2019), MMCNN (Wang et al. 2019b), MEMC-Net (Bao et al. 2021), RRCN (Li et al. 2019a), RTVSR (Bare et al. 2019), MultiBoot VSR (Kalarot and Porikli 2019), TecoGAN (Chu et al. 2020), MuCAN (Li et al. 2020), BasicVSR (Chan et al. 2021d). DC: EDVR (Wang et al. 2019c), DNLN (Wang et al. 2019a), TDAN (Tian et al. 2020), D3Dnet (Ying et al. 2020), VESR-Net (Chen et al. 2020). 2D Conv:VSRResFeatGAN (Lucas et al. 2019), FFCVSR (Yan et al. 2019). 3D Conv: DUF (Jo et al. 2018), FSTRN (Li et al. 2019b), 3DSRnet (Kim et al. 2019), DSMC (Liu et al. 2021b). RCNN: BRCN (Huang et al. 2015, 2018), STCN (Guo and Chao 2017), RISTN (Zhu et al. 2019b), RLSP (Fuoli et al. 2019a), RSDN (Isobe et al. 2020). Non-Local: PFNL (Yi et al. 2019). Other: RBPN (Haris et al. 2019), STARnet (Haris et al. 2020), DNSTNet (Sun et al. 2020)

sliding-window, while the subsequent methods resort to longer sequences. For example, BasicVSR employs bidirectional RCNN, by which the features are propagated forward and backward independently. Moreover, the non-local subnetwork, such as in the PFNL method, aims to compute the correlations between all possible pixels within and across frames. These indicate that the methods tend to capture longer-range dependencies in the video sequences, and they expand the receptive field in the network from local to global. (3) In the MEMC methods such as FRVSR, STTN, SOFVSR, TecoGAN, and MuCAN, most of them adopt deep learning techniques for estimation the optical flow, since the deep learning may have adaptive ability for various data than the conventional methods. (4) The practicality of methods. As the requirements for super-resolving of higher quality videos develop, the recently proposed methods also become more practicable. The test videos evolve from Vid4 and UVGD, to REDS. All the discussions indicate that we will mainly focus on the methods for the videos with more complex motions and scene changes.

4 Methods with alignment

The methods with alignment make neighboring frames explicitly align with the target frame by using extracted motion information before subsequent reconstruction. These methods mainly use motion estimation and motion compensation (MEMC) or deformable convolution, which are two common techniques for aligning frames. Next we will introduce state-of-the-art methods based on each of the techniques in detail.

4.1 Motion estimation and compensation methods

In the methods with alignment for video super-resolution, most of them apply the motion estimation and motion compensation techniques. Specifically, the purpose of motion estimation is to extract inter-frame motion information, while motion compensation is used to perform the warping operation between frames according to inter-frame motion information and to make one frame align with another frame. A majority of the motion estimation techniques are performed by the optical flow method (Dosovitskiy et al. 2015). This method tries to calculate the motion between two neighboring frames through their correlations and variations in the temporal domain. The motion estimation methods can be divided into two categories: traditional methods (e.g., Lucas and Kanade (1981); Drulea and Nedevschi (2011)) and deep learning methods such as FlowNet (Dosovitskiy et al. 2015), FlowNet 2.0 (Ilg et al. 2017) and SpyNet (Ranjan and Black 2017).

In general, an optical flow method takes two frames (e.g., I_i and I_j) as inputs. One is the target frame and the other is the neighboring frame. Then the method computes a vector field of optical flow $F_{i\to j}$ from the frame I_i to I_j by the following formula:

$$F_{i\to j}(h_{i\to j}, v_{i\to j}) = ME(I_i, I_j; \theta_{ME}) \tag{7}$$

where $h_{i\to j}$ and $v_{i\to j}$ is the horizontal and vertical components of $F_{i\to j}$, $ME(\cdot)$ is a function used to compute optical flow, and θ_{ME} is its parameter.

The motion compensation is used to perform image transformation between images in terms of motion information to make neighboring frames align with the target frame. In general, a compensated frame I'_j is expressed as:

$$I'_j = MC(I_i, F_{i\to j}; \theta_{MC}) \tag{8}$$

where $MC(\cdot)$ is a motion compensation function, I_i, $F_{i\to j}$ and θ_{MC} are the neighboring frame, optical flow and the parameter. MC can be achieved by some methods such as bilinear interpolation and spatial transformer network (STN) (Jaderberg et al. 2015). An example of motion estimation and motion compensation is shown in Fig. 3.

Both the ME and MC processes can be conducted by a deep learning method or traditional one (non-deep learning). According to the technique that applies to ME or MC is traditional or deep learning, we further divide the MEMC methods into two subcategories. If any of the processes in ME or MC utilizes a deep neural network, then the method falls into the deep learning category, otherwise the method belongs to the traditional one. Therefore, the traditional methods in the MEMC methods comprise of the following three ones: Deep-DE (Liao et al. 2015), VSRNet (Kappeler et al. 2016), and RRCN (Li et al. 2019a). The other MEMC methods are included in the deep learning subcategory. Below we depict some representative methods in detail.

4.1.1 Deep-DE

The deep draft-ensemble learning method (Deep-DE)[1] (Liao et al. 2015) has two phases, as shown in Fig. 4. It first generates a series of SR drafts by adjusting the TV-ℓ_1 flow (Brox et al. 2004; Guo and Chao 2017) and the motion detail preserving (MDP) (Xu et al. 2012).

[1] Code: http://www.cse.cuhk.edu.hk/leojia/projects/DeepSR/.

Then both the SR drafts and the bicubic-interpolated LR target frame are fed into a CNN for feature extraction, fusion and super-resolution.

The CNN in Deep-DE consists of four convolutional layers: the first three layers are general convolutional layers, and the last layer is a deconvolution layer. Their kernel sizes are 11×11, 1×1, 3×3 and 25×25, respectively, and the numbers of channels are 256, 512, 1 and 1.

4.1.2 VSRnet

VSRnet[2] (Kappeler et al. 2016) is based on the image super-resolution method, SRCNN (Dong et al. 2014), and its network architecture is shown in Fig. 5. VSRnet mainly consists of motion estimation and compensation modules, and three convolutional layers, and each convolutional layer is followed by a rectified linear unit (ReLU) except for the last one. The main difference between VSRnet and SRCNN is the number of input frames. That is, SRCNN takes a single frame as input, while VSRnet uses multiple successive frames, which are compensated frames. The motion information between frames is computed by the Druleas algorithm (Drulea and Nedevschi 2011). In addition, VSRnet proposes a filter symmetry enforcement (FSE) mechanism and an adaptive motion compensation mechanism, which are separately used to accelerate training and reduce the impact of unreliable compensated frames, and thus can improve video super-resolution performance.

4.1.3 RRCN

The residual recurrent convolutional network (RRCN) (Li et al. 2019a), as shown in Fig. 6, is a bidirectional recurrent neural network, which learns a residual image. RRCN proposes an unsynchronized full recurrent convolutional network, where unsynchronization refers to the input of multiple consecutive video frames, and only the middle one is super-resolved.

RRCN uses the combined local-global with total variable (GLG-TV) method (Drulea and Nedevschi 2011) to perform motion estimation and compensation for the target frame and its adjacent frames. The compensated frames are used as input to the network. The forward convolution and recurrent convolution are conducted in the forward network and the backward network, respectively, and their outputs are summed up. Finally, the result is obtained by adding the target frame to the input. In order to further improve the performance, RRCN also uses the self-ensemble strategy and combines it with the output of the single image super-resolution method, EDSR+ (Lim et al. 2017), to obtain two models named RRCN+ and RRCN++, respectively.

4.1.4 VESPCN

The video efficient sub-pixel convolutional network (VESPCN) (Caballero et al. 2017) proposes a spatial motion compensation transformer (MCT) module for motion estimation and compensation. Then the compensated frames are fed into a series of convolutional layers for feature extraction and fusion, as shown in Fig. 7. Finally, the super-resolution results are obtained through a sub-pixel convolutional layer for upsampling.

[2] Code: https://superresolution.tf.fau.de/.

(a) A target frame (b) Its neighboring frame

(c) The compensated image (d) The estimated optical flow image

Fig. 3 An example of motion estimation and compensation. Note that the small rightmost image is the legend of (d). Different colors represent different directions of motion and the intensity of the color is the range of motion. (Color figure online)

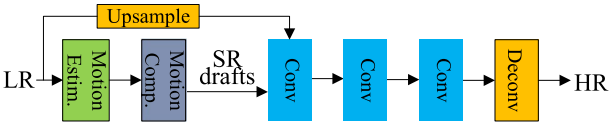

Fig. 4 The architecture of Deep-DE (Liao et al. 2015). Here Motion Estim. is a motion estimation block, Motion Comp. is a motion compensation block, Conv is a convolutional layer and Deconv is a deconvolutional layer

The MCT module adopts CNNs to extract motion information and perform motion compensation. MCT uses a coarse-to-fine approach to compute the optical flow for image sequences. Firstly, in the coarse estimation stage, the network takes two consecutive frames (i.e., the target frame and a neighboring frame) as inputs. The coarse network consists of 5 convolutional layers and a sub-pixel convolutional layer. And it first performs the $\times 2$ downsampling operation two times and then performs the $\times 4$ upsampling operation by a sub-pixel convolutional layer to get coarse optical flow estimation results. Secondly, the neighboring frame is warped according to the optical flow. In the fine estimation stage, the target frame, neighboring frame, optical flow computed in the coarse stage and the warped neighboring frame are the input of the fine network, whose architecture is similar to the coarse network. It first conducts $\times 2$ downsampling and then perform $\times 2$ upsampling at the end of the network to attain the fine optical flow. Together with the coarse optical flow, the fine optical flow is used to obtain the final estimation result. Finally, the neighboring frame is warped again by the final optical flow to make the warped frame align with the target frame.

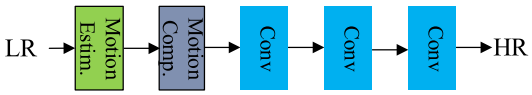

Fig. 5 The network architecture of VSRnet (Kappeler et al. 2016)

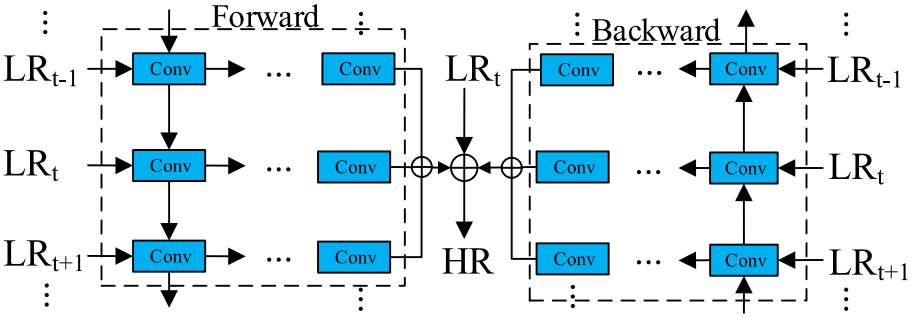

Fig. 6 The network architecture of RRCN (Li et al. 2019a)

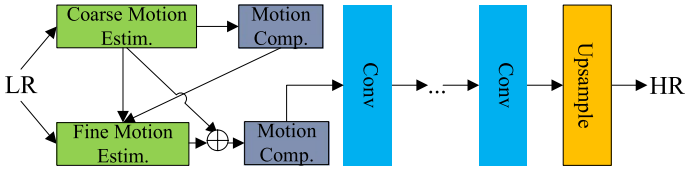

Fig. 7 The network architecture of VESPCN (Caballero et al. 2017). Here ⊕ denotes element-wise sum

4.1.5 DRVSR

The detail-revealing deep video super-resolution (DRVSR)[3] (Tao et al. 2017) method proposes a sub-pixel motion compensation layer (SPMC) that can perform the up-sampling and motion compensation operations simultaneously for neighboring input frames according to the estimated optical flow information. The network architecture of DRVSR is illustrated in Fig. 8.

DRVSR consists of three main modules: a motion estimation module, a motion compensation module using the SPMC layer, and a fusion module. The motion estimation module is implemented by the motion compensation transformer (MCT) network (Caballero et al. 2017). The SPMC layer consists of two sub-modules, namely grid generator and sampler. The grid generator first transforms the coordinates in the LR space into the coordinates in the HR space according to the optical flow, and then the sampler performs the interpolation operation in the HR space. In the fusion module, it applies the convolution with stride 2 to perform down-sampling and then conducts the deconvolution for up-sampling to obtain the HR residual image of the target frame. Together with the upsampled LR target frame,

[3] Code: https://github.com/jiangsutx/SPMC_VideoSR.

this residual image yields the final result. DRVSR also adopts the ConvLSTM module (Shi et al. 2015) to handle spatio-temporal information.

4.1.6 RVSR

Robust video super-resolution (RVSR) (Liu et al. 2017) proposes a spatial alignment module to attain great alignment performance and a temporal adaptive module to adaptively determine the optimal scale of temporal dependency. And its architecture is shown in Fig. 9.

The spatial alignment module is responsible for the alignment of the multi-frames so that the neighboring frames are aligned with the target frame. It first estimates the transformation parameters between the neighboring frame and the target frame through a localization net, and then makes the neighboring frame align with the target frame through a spatial transformation layer (Jaderberg et al. 2015) based on the obtained parameters. The localization net consists of two convolutional layers, each of which is followed by a max-pooling layer, and two fully connected layers. The temporal adaptive module is composed of multiple branches of SR subnetwork and a temporal modulation. Each subnetwork is responsible for handling a temporal scale (i.e., the number of input frames), and outputting the corresponding super-resolution result. Then the super-resolution result of each subnetwork is allocated a weight through the temporal modulation. The final super-resolution result is the weight sum of the super-resolution result of each branch and its weight. The number of the input frames of the temporal modulation module is identical to the maximum number of input frames in the super-resolution network, and the network structure of the temporal modulation module is the same as that of the super-resolution network, and both of them are based on the structure of ESPCN (Shi et al. 2016).

4.1.7 FRVSR

Frame recurrent video super-resolution (FRVSR)[4] (Sajjadi et al. 2018) mainly proposes to use the previously inferred HR estimate to super-resolve the subsequent frame for producing temporally consistent results and reducing computational cost. The architecture of FRVSR is illustrated in Fig. 10.

The detailed implementation adopts an optical estimation network to compute the optical flow from the previous frame to the target frame. Then the LR optical flow is upsampled to the same size with the HR video by bilinear interpolation. The HR variant of the previous frame is warped by the upsampled LR optical flow, and then the warped HR frame is downsampled by space-to-depth transformation to get the LR version. Finally, the LR variant of the warped HR frame and the target frame are fed into the subsequent super-resolution network to attain the result for the target frame. In FRVSR, the optical flow network consists of 14 convolutional layers, 3 pooling layers and 3 bilinear upsampling layers. Each convolutional layer is followed by a LeakyReLU activation function, except for the last convolutional layer. The super-resolution network consists of 2 convolutional layers, 2 deconvolution layers with $\times 2$ and 10 residual blocks, where each residual block consists of 2 convolutional layers and a ReLU activation function.

[4] Code: https://github.com/msmsajjadi/FRVSR.

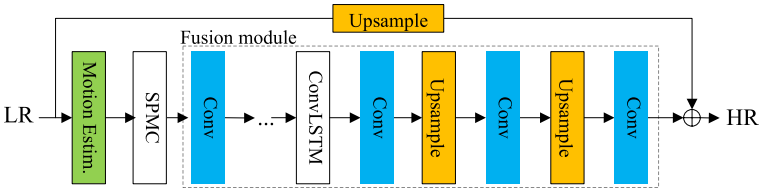

Fig. 8 The network architecture of DRVSR (Tao et al. 2017). Here SPMC denotes a sub-pixel motion compensation layer, ConvLSTM is the convolutional LSTM (Shi et al. 2015)

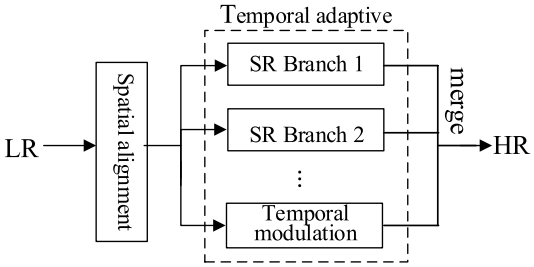

Fig. 9 The network architecture of RVSR (Liu et al. 2017), where SR denotes Super-Resolution

4.1.8 STTN

Spatio-temporal transformer network (STTN) (Kim et al. 2018a) proposes a spatio-temporal transformer module, which is used to address the problem that previous optical flow methods only process a pair of video frames, which may cause inaccurate estimation when occlusion and luminance variation exist in videos. The proposed module can handle multiple frames at a time. The architecture of STTN is illustrated in Fig. 11.

STTN consists of three major modules: a spatio-temporal flow estimation module, a spatio-temporal sampler module, and a super-resolution module. The first module is a U-style network, similar to U-Net (Ronneberger et al. 2015), consisting of 12 convolutional layers and two up-sampling layers. It first performs ×4 downsampling, and then ×4 up-sampling to restore the size of the input frames. This module is responsible for optical flow estimation of the consecutive input frames including the target frame and multiple neighboring frames, and the final output is a 3-channel spatio-temporal flow that expresses the spatial and temporal changes between frames. The spatio-temporal sampler module is actually a trilinear interpolation method, which is responsible for performing warp operation for current multiple neighboring frames and obtaining the aligned video frames according to the spatio-temporal flow obtained by the spatio-temporal flow module. For video super-resolution, the aligned frames can then be fed into the super-resolution (SR) module for feature fusion and super-resolution of the target frame.

4.1.9 SOFVSR

Super-resolution optical flow for video super-resolution (SOFVSR)[5] (Wang et al. 2019d) is proposed to super-resolve LR estimated optical flow for attaining great SR performance, and its architecture is shown in Fig. 12.

[5] Code: https://github.com/LongguangWang/SOF-VSR.

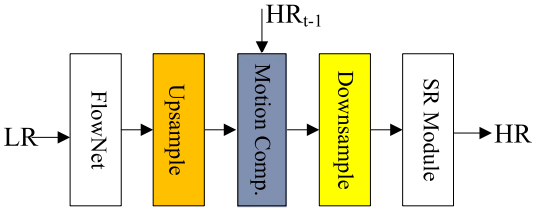

Fig. 10 The network architecture of FRVSR (Sajjadi et al. 2018). Here FlowNet is an optical flow estimation module, and SR Module is a super-resolution module

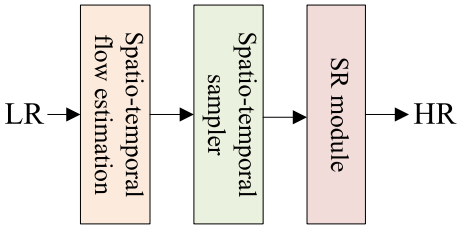

Fig. 11 The network architecture of STTN (Kim et al. 2018a)

The optical flow between frames is estimated by a coarse-to-fine approach including the optical flow reconstruction network (OFRnet), which finally yields a high-resolution optical flow. Then the HR optical flow is converted to the LR optical flow by a space-to-depth transformation. The neighboring frames are warped by the LR optical flow to make the neighboring frames align with the target frame. Then the super-resolution network (SRnet) takes the target frame and warped frames as inputs to obtain the final super-resolution result. SRnet consists of two convolutional layers, five residual dense blocks and a sub-pixel convolutional layer.

4.1.10 TOFlow

The architecture of the task-oriented flow (TOFlow)[6] (Xue et al. 2019) is shown in Fig. 13. TOFlow combines the network for optical flow estimation with the reconstruction network, and trains them jointly to obtain optical flow network tailored to a specific task such as video SR, video interpolation and video deblurring.

TOFlow adopts SpyNet (Ranjan and Black 2017) as the network for the optical flow estimation, and then adopts a spatial transformer network (STN) to warp the neighboring frame according to the computed optical flow. Then the final result is obtained by an image processing network. For the video super-resolution task, the image processing module consists of 4 convolutional layers, where kernel sizes are 9×9, 9×9, 1×1, and 1×1, respectively, and the numbers of channels are 64, 64, 64, and 3, respectively.

[6] Code: https://github.com/anchen1011/toflow.

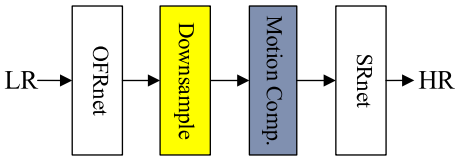

Fig. 12 The network architecture of SOFVSR (Wang et al. 2019d). Here, OFRnet is an optical flow network, and SRnet is a super-resolution module

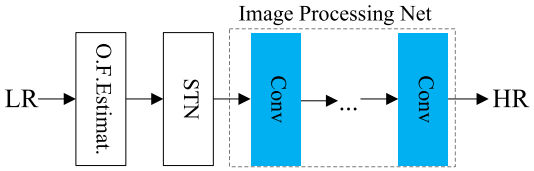

Fig. 13 The network architecture of TOFlow (Xue et al. 2019). Here O.F.Estimat. is the optical flow estimation, STN is a spatial transformer network

4.1.11 MMCNN

The architecture of the multi-memory convolutional neural network (MMCNN)[7] (Wang et al. 2019b) is shown in Fig. 14, and it consists of 5 major modules: optical flow module for motion estimation and motion compensation, feature extraction, multi-memory detail fusion, feature reconstruction, and upsample modules, where the last module uses a sub-pixel convolutional layer.

Consecutive input frames are first processed by the optical flow estimation module to make neighboring frames align with the target frame and then the warped frames are fed into subsequent network modules to attain the residual image of the target frame. Finally, this residual image is added into the upsampled LR target frame, which is computed by bicubic interpolation, to obtain the super-resolution result. In the multi-memory detail fusion module, MMCNN adopts the ConvLSTM module (Shi et al. 2015) to merge the spatio-temporal information. Moreover, the feature extraction, detail fusion, and feature reconstruction modules are all built based on residual dense blocks (Zhang et al. 2018c; Huang et al. 2017), where the key difference among them is merely the type of network layers.

4.1.12 MEMC-Net

The motion estimation and motion compensation network (MEMC-Net)[8] (Bao et al. 2021), as shown in Fig. 15, mainly proposes an adaptive warping layer.

The adaptive warping layer warps the neighboring frame through the estimated optical flow and the convolutional kernel, which are resulted from a motion estimation network and a kernel estimation network, respectively, and aligns the neighboring frame with the target frame. The motion estimation network adopts FlowNet (Dosovitskiy et al. 2015), and the kernel estimation network uses an improved U-Net (Ronneberger et al. 2015) including five max-pooling layers, five un-pooling layers and skip connections from the encoder to the decoder. In MEMC-Net, the architecture of the super-resolution module, namely frame enhancement module, is similar to that of EDSR (Lim et al. 2017). In order to deal with the

[7] Code: https://github.com/psychopa4/MMCNN.
[8] Code: https://github.com/baowenbo/MEMC-Net.

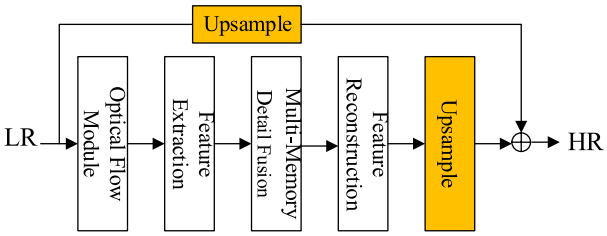

Fig. 14 The network architecture of MMCNN (Wang et al. 2019b)

occlusion problem, it adopts a pre-trained ResNet18 (He et al. 2016) to extract the feature of input frames. Moreover, it feeds the output of the first convolutional layer of ResNet18 as the context information into the adaptive warping layer to perform the same operation.

4.1.13 RTVSR

The real-time video super-resolution (RTVSR) (Bare et al. 2019), as shown in Fig. 16, adopts a convolutional network called motion convolutional kernel estimation network, which is a full convolution codec structure, to estimate the motion between the target frame and the neighboring frame and produce a pair of 1D convolutional kernel corresponding to the current target frame and neighboring frame. Then the neighboring frame is warped by using estimated convolutional kernels to make it align with the target frame. RTVSR designs an important component called gated enhance units (GEUs) to learn useful features, which is an improved variant based on Li et al. (2018).

4.1.14 MultiBoot VSR

The multi-stage multi-reference bootstrapping for video super-resolution (MultiBoot VSR) (Kalarot and Porikli 2019) consists of two stages. That is, in order to further improve performance, the output of the first stage is used as the input of the second stage. The network architecture of MultiBoot VSR is shown in Fig. 17.

The LR frames are input to the FlowNet 2.0 to compute optical flow and perform the motion compensation. Then the processed frames are fed into the first-stage network to attain the super-resolution result of the target frame. In the second stage of MultiBoot VSR, the output from the previous stage is downsampled, concatenated with the initial LR frame, and then input to the network to obtain final super-resolution result for the target frame.

4.1.15 MuCAN

The architecture of multi-correspondence aggregation network (MuCAN) (Li et al. 2020) is shown in Fig. 18. MuCAN is an end-to-end network consisting of a temporal multi-correspondence aggregation module (TM-CAM), a cross-scale non-local-correspondence aggregation module (CN-CAM), and a reconstruction module.

In TM-CAM, two neighboring LR frames are first encoded into lower-resolution features to be more stable and robust to noise. Then the aggregation starts in the original LR feature space by an aggregation unit (AU) to compensate large motion while progressively moving up to low-level/high-resolution stages for subtle sub-pixel shift. In a single AU, a patch-based matching strategy is used since it naturally contains structural information.

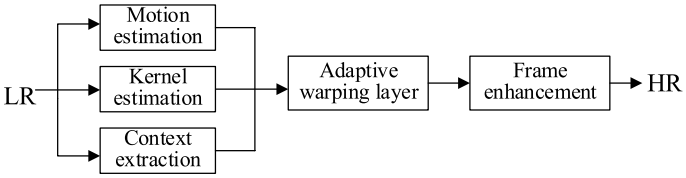

Fig. 15 The network architecture of MEMC-Net (Bao et al. 2021)

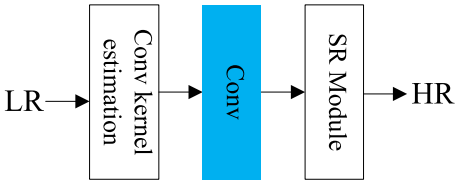

Fig. 16 The network architecture of RTVSR (Bare et al. 2019). Here SR Module denotes super-resolution module

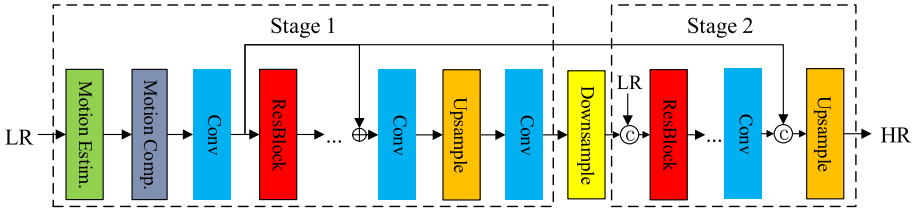

Fig. 17 The network architecture of MultiBoot VSR (Kalarot and Porikli 2019)

Multiple candidates are then aggregated to obtain sufficient context information. The aggregated information is then passed to CN-CAM, which then uses a pyramid structure based on AvgPool to execute spatio-temporal non-local attention and coarse-to-fine spatial attention. Finally, the results are aggregated and sent to the reconstruction module to yield the final HR result.

4.1.16 TecoGAN

Temporally coherent GAN (TecoGAN)[9] (Chu et al. 2020) mainly proposes a spatio-temporal discriminator for realistic and coherent video super-resolution, and a novel "Ping-Pong" loss to tackle recurrent artifacts. Like GAN, TecoGAN also consists of a generator and a discriminator and its architecture is shown in Fig. 19.

The generator takes the target frame, the previous frame and previous estimated HR frames as inputs. First, input frames are fed into the optical flow module, which is a CNN similar to the optical flow estimation module in FRVSR (Sajjadi et al. 2018). In this module, the LR optical flow between the target frame and neighboring frames is estimated and enlarged by the bicubic interpolation to attain the corresponding HR optical flow. Then the previous HR frame is warped by the HR optical flow. The warped previous HR frame and target frame are fed into subsequent convolutional modules that include two convolutional

[9] Code: https://github.com/thunil/TecoGAN.

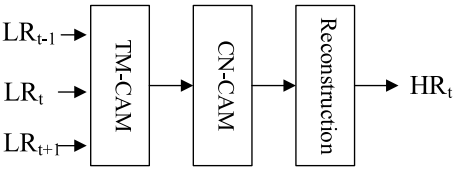

Fig. 18 The network architecture of MuCAN (Li et al. 2020)

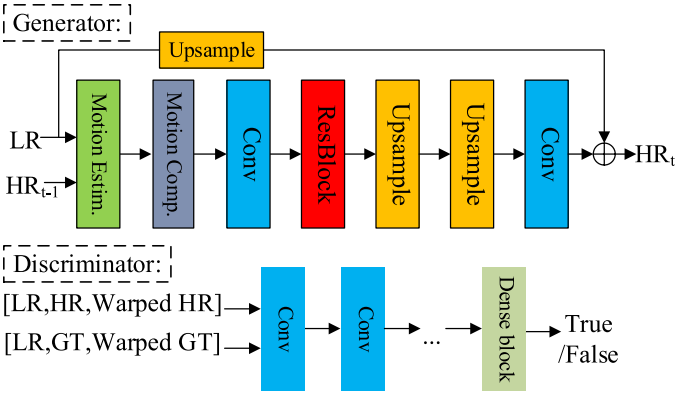

Fig. 19 The network architecture of TecoGAN (Chu et al. 2020)

layers, a residual block and two upsample modules with a deconvolution layer, to yield a restored target frame. Moreover, the discriminator assesses the quality of super-resolution results. The discriminator takes the generated results and GT as inputs, where each of them has three components, that is, three consecutive HR frames, three corresponding upsampled LR frames and three warped HR frames. With such input formats, the spatial over-smooth and temporal inconsistence in the final results can be relieved. TecoGAN also proposes a "ping-pong" loss function to reduce the long-term temporal detail drift and make super-resolution results more natural.

4.1.17 BasicVSR

The authors proposed a generic framework for video super resolution, called BasicVSR, as shown in Fig. 20. It is a typical bidirectional recurrent network, which mainly consists of three modules: the backward (B) module, the forward (F) module, and the upsampling (U) module. The B module receives the output of the next B module, current frame, and the following frame, while the F module receives the output of the previous F module, current frame, and the preceding frame. Then the outputs of the two modules are fused through a U module to yield the super-resolved current frame. These processes iterate until all the frames are super-resolved. The B/F module composes of generic components: the motion estimation, spatial warping, and residual blocks. The authors further propose two processing mechanisms the information-refill and coupled propagation, which consist of the IconVSR algorithm. The former addresses the performance degradation caused by misalignment, and the latter deals with the lack of information interaction between the forward processing and the backward processing in BasicVSR. In the information-refill mechanism,

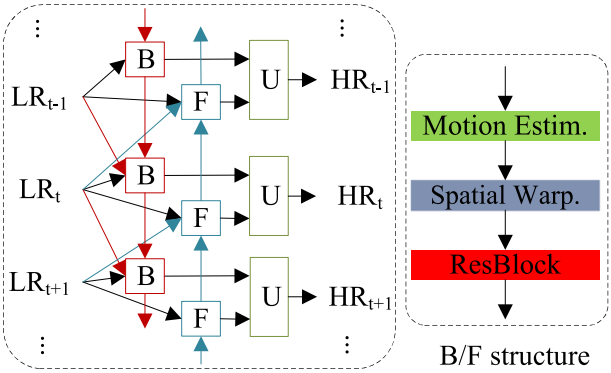

Fig. 20 The network architecture of BasicVSR (Chan et al. 2021d)

if the currently processed frame is in the selected keyframe set, it will be fused; otherwise, the aligned result will be directly sent into the residual block without fusion. This mechanism relieves error accumulation caused by misalignment, thus avoiding the performance degradation. In the coupling propagation mechanism, the output of backward propagation is directly used as the input of forward propagation, so as to achieve information interaction between them.

In summary, the MEMC techniques are used to align neighboring frames with a target frame, and are probably the most common method for solving video super-resolution tasks. However, the problem is that they cannot guarantee the accuracy of motion estimation when lighting changes dramatically or there are large motions in videos. In these cases, the performance of the video super-resolution degrades greatly. This is confirmed by the assumption in Lucas and Kanade (1981). When dealing with complex motions (not only large motions) and varying illumination, the calculation of motion estimation based on optical flow methods may break the hypothesis of brightness consistency, small moti on, and spatial coherence. Then the estimation of optical flow becomes inaccurate, and there arises errors, which easily results in artifacts and blurring. To address this issue, the methods with alignment (e.g., the deformable convolution which is presented as one module in the deep network to align frames) and the methods without alignment are both proposed.

4.2 Deformable convolution methods

The deformable convolutional network was first proposed by Dai et al. (2017) and the improved variant (Zhu et al. 2019a) was proposed in 2019. In ordinary CNNs, the convention is to use a fixed geometric structure in a layer, which restricts the network's capability to model geometric transformations. In contrast, the deformable convolution is able to overcome this limitation. The illustration of the deformable convolution for feature alignment is shown in Fig. 21. The target feature maps concatenating with the neighboring feature maps are projected to attain offsets via additional convolutional layers. The offsets are applied to the conventional convolution kernel to yield a deformable convolution kernel, and then it is convolved with the input feature maps to produce the output feature maps. The methods that adopt deformable convolution mainly include the enhanced deformable video restoration (EDVR) (Wang et al. 2019c), deformable non-local network (DNLN) (Wang et al. 2019a), and temporally deformable alignment network (TDAN) (Tian et al. 2020), which are depicted in detail as follows.

4.2.1 EDVR

The enhanced deformable video restoration (EDVR)[10] (Wang et al. 2019c), as shown in Fig. 22, is the champion model in the NTIRE19 Challenge (Nah et al. 2019a, b). EDVR proposes two key modules: the pyramid, cascading and deformable (PCD) alignment module as in Ranjan and Black (2017), Sun et al. (2018), Hui et al. (2018, 2021a) and the temporal-spatial attention (TSA) fusion module, which are used to solve large motions in videos and to effectively fuse multiple frames, respectively.

EDVR mainly consists of four parts: one PCD alignment module, a TSA fusion module, a reconstruction module, and an upsample module using a sub-pixel convolutional layer. Firstly, the input frames are aligned by the PCD alignment module, and then the aligned frames are fused by the TSA fusion module. Then the fused results are fed into the reconstruction module to refine the features, and then through the up-sampling, a HR image called the residual image is obtained. The final result is obtained by adding the residual image to a direct upsampling target frame. To further improve performance, EDVR also adopts a two-phase approach, whose second phase is similar to the first but with a shallower network depth.

4.2.2 DNLN

The deformable non-local network (DNLN)[11] (Wang et al. 2019a), as shown in Fig. 23, designs an alignment module and a non-local attention module based on the deformable convolution (Dai et al. 2017; Zhu et al. 2019a) and non-local networks (Wang et al. 2018), respectively. The alignment module uses the hierarchical feature fusion module (HFFB) (Hui et al. 2021b) within the original deformable convolution to generate convolutional parameters. Moreover, DNLN utilizes multiple deformable convolutions in a cascaded way, which makes inter-frame alignment more accurate.

4.2.3 TDAN

The temporally deformable alignment network (TDAN)[12] (Tian et al. 2020), as shown in Fig. 24, applies deformable convolution to the target frame and the neighboring frame, and attains corresponding offsets. Then the neighboring frame is warped in terms of the offsets to align with the target frame. TDAN is divided into three parts, i.e., a feature extraction module, a deformable convolution module and a reconstruction module.

4.2.4 D3Dnet

The architecture of the deformable 3D convolution network (D3Dnet)[13] (Ying et al. 2020) is shown in Fig. 25. D3Dnet proposes 3D deformable convolution to achieve strong spatio-temporal feature modeling capability. The inputs are first fed to a 3D convolutional layer to

[10] Code: https://github.com/xinntao/EDVR.
[11] Code: https://github.com/wh1h/DNLN.
[12] Code: https://github.com/YapengTian/TDAN-VSR-CVPR-2020.
[13] Code: https://github.com/XinyiYing/D3Dnet.

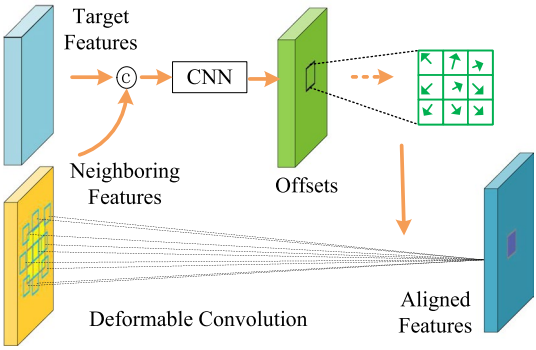

Fig. 21 Deformable convolution for frame alignment

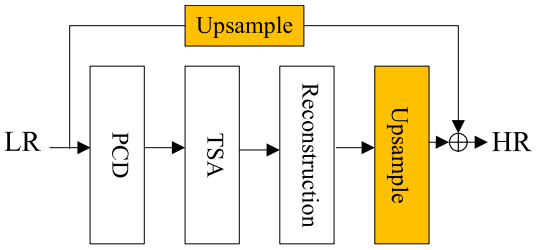

Fig. 22 The network architecture of EDVR (Wang et al. 2019c), where PCD is the pyramid, cascading and deformable alignment module, and TSA is the temporal-spatial attention fusion module

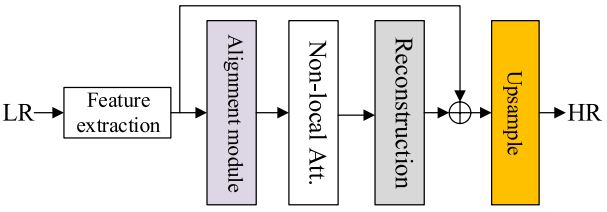

Fig. 23 The network architecture of DNLN (Wang et al. 2019a). Here Non-local Att. is the non-local attention module

generate features, which are then fed to 5 Residual Deformable 3D Convolution (ResD3D) blocks to achieve motion compensation and capture spatial information.

4.2.5 VESR-Net

The architecture of video enhancement and super-resolution network (VESR-Net) (Chen et al. 2020), as shown in Fig. 26, is the champion model in the Youku video enhancement and super-resolution challenge. VESR-Net mainly consists of a feature encoder, a fusion module and a reconstruction module.

The LR frames are firstly processed by the feature encoder consisting of a convolution layer and several stacked channel-attention residual blocks (CARBs) (Zhang et al. 2018b). Then in the fusion module, the PCD convolution in Wang et al. (2019c) performs the inter-frame feature alignment. The separate non-local submodule (Separate NL) divides feature maps in spatial, channel and temporal dimensions and processes them to obtain correlation information separately. In contrast to the vanilla non-local (Wang et al. 2018) architecture, Separate NL can fuse the information across video frames and across pixels in each frame with less parameters and shallower network. Finally, VESR-Net utilizes CARBs followed

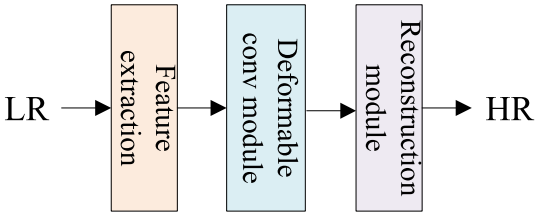

Fig. 24 The network architecture of TDAN (Tian et al. 2020)

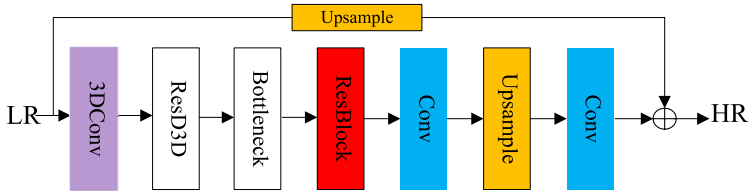

Fig. 25 The network architecture of D3Dnet (Ying et al. 2020)

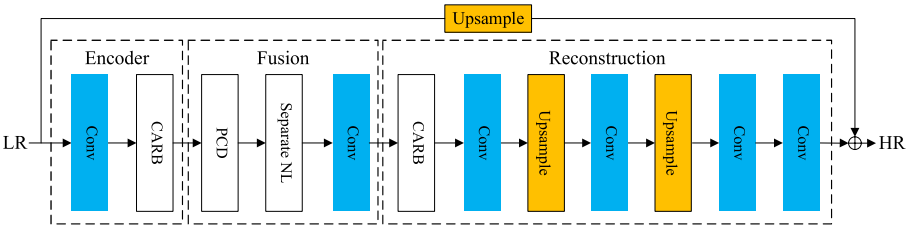

Fig. 26 The architecture of VESR-Net (Chen et al. 2020), where CARB is the channel-attention residual block, and Separate NL denotes the separate non-local architecture

with a feature decoder for upsampling in the reconstruction module, where the upsample module is implemented by a sub-pixel convolutional layer. And it outputs the super-resolved frame by adding with the bicubic-interpolation LR target frame.

The evolution of methods with alignment. In the methods with alignment, the motion estimation and motion compensation techniques, as a classic research topic in computer vision, have been applied to video super-resolution in the early years. MEMC has wide range of applications such as video coding and enhancing the interlaced scanning. As the advent of deep learning based VSR, many works employ MEMC to capture the motion information contained in video frames. The early work of MEMC is Deep-DE (Liao et al. 2015), and some recently proposed methods such as VESPCN (Caballero et al. 2017), SOFVSR (Wang et al. 2019d), TOFlow (Xue et al. 2019) and FRVSR (Sajjadi et al. 2018) also adopted MEMC techniques. Specifically, early video super-resolution algorithms adopt traditional MEMC methods such as Druleas in VSRnet (Kappeler et al. 2016), while subsequent algorithms such as VESPCN (Caballero et al. 2017), TOFlow (Xue et al. 2019) and FRVSR (Sajjadi et al. 2018) mainly design sub-module or sub-network for MEMC.

However, the accuracy of most MEMC methods is usually not guaranteed. When the luminance changes or the videos contain large motions between frames, the performance of VSR degrades dramatically. Hence, the deformable convolution (DConv), which is

not sensitive to varying lighting and motion conditions, has attracted more attention from researchers. DConv applies a learnable offset to each sampling point compared with the conventional convolution. Therefore, DConv can not only expand the receptive field of convolution kernel, but also enrich the shape of receptive field. When handling varying lighting and motion conditions, the conventional convolution with fixed kernel and limited receptive field may not be capable of capturing varying conditions. While DConv uses a learnable parameter for the kernel to analyze lighting and motion features, which can better capture complex motions and illumination changes. The deformable convolution was proposed by Dai et al. (2017) to enhance the transformation modeling capability of CNNs for the geometric variations of objects. In the VSR methods, TDAN (Tian et al. 2020) first utilized it to perform inter-frame alignment. After that, DNLN (Wang et al. 2019a), EDVR (Wang et al. 2019c), and D3Dnet (Ying et al. 2020) further promote it for frame alignment. Nevertheless, the deformable convolution still has some drawbacks including high computational complexity and harsh convergence conditions. Therefore, there is still room for improvement of this technique in the future.

In addition, the performance of the MEMC-based methods will degrade greatly when there were dramatic lighting changes and large motions in videos. Nevertheless, the network architecture is one of the important factors to affect its performance. Other factors include the training dataset, training strategy, data preprocessing, hyper-parameter setting, iteration times, etc. Although the MEMC-based methods have the limitation to deal with videos containing lighting changes and large motions, they can be counteracted by other network designs and training settings. For example, BasicVSR/IconVSR adopts a bidirectional recurrent network as backbone, which fully utilizes the global information from the video sequences and expands receptive field. Thus, they may gain superior performance compared with the other MEMC methods, which mainly use convolutions. Moreover, the training process, which uses a Cosine annealing scheme (Loshchilov and Hutter 2017), is probably more refined.

5 Methods without alignment

In contrast to the methods with alignment, the methods without alignment do not align neighboring frames for video super-resolution. This type of methods mainly exploit the spatial or spatio-temporal information for feature extraction. According to the dominating techniques utilized for initial feature extraction, we further categorize them into five types: the 2D convolution methods (2D Conv), 3D convolution methods (3D Conv), recurrent convolutional neural network (RCNN), non-local network based, and other methods. Among them, the first type falls into the spatial methods, while the following three are the spatio-temporal methods, whose characteristic is to exploit both the spatial and temporal information from input videos. Other methods include the ones do not belong to any of the former. We present them in detail as follows.

5.1 2D convolution methods

Instead of alignment operations such as motion estimation and motion compensation between frames, the input frames are directly fed into a 2D convolutional network to spatially perform feature extraction, fusion and super-resolution operations. This may be a simple approach for solving the video super-resolution problem since it makes the network

learn the correlation information within frames by itself. The representative methods are VSRResFeatGAN (Lucas et al. 2019) and FFCVSR (Yan et al. 2019).

5.1.1 VSRResFeatGAN

VSRResFeatGAN (Lucas et al. 2019) utilizes GAN to address VSR tasks and find a good solution by adversarial training. The generator shown in Fig. 27 consists of convolutional layers and residual blocks. And each residual block is composed of two convolutional layers and is followed by a ReLU activation function. Moreover, the discriminator consists of three groups of convolutions and a fully connected layer, where each group includes a convolutional layer, Batch Normalization (BN), and LeakyReLU. The discriminator determines whether the output of the generator is a generated image or GT image. Then the result of the discriminator reacts to the generator, and promotes it to yield results closer to the GT images. Finally, a relative satisfactory solution is obtained through an iterative optimization.

5.1.2 FFCVSR

The architecture of the frame and feature-context video super-resolution (FFCVSR)[14] (Yan et al. 2019) is shown in Fig. 28. Unlike common MEMC techniques, FFCVSR consists of several local networks and context networks and utilizes inter-frame information in a different way. The LR unaligned video frames and the HR output of the previous frame are directly taken as inputs to the network for the purpose of restoring high-frequency details and maintaining temporal consistency.

In summary, the above two methods both exploit spatial correlation between frames for VSR tasks. VSRResFeatGAN utilizes adversarial training of GANs to find an appropriate solution. As the discriminator in GANs has to guess whether the generated frame is close to the ground truth, the VSR results in terms of PSNR and SSIM are not always satisfactory compared with other methods, such as FFCVSR.

5.2 3D convolution methods

The 3D convolutional module (Tran et al. 2015; Ji et al. 2013) operates on spatio-temporal domain, compared with 2D convolution, which only utilizes spatial information through the sliding kernel over input frame. This is beneficial to the processing of video sequences, as the correlations among frames are considered by extracting temporal information. The representative 3D convolution methods for VSR are DUF (Jo et al. 2018), FSTRN (Li et al. 2019b), and 3DSRnet (Kim et al. 2019).

5.2.1 DUF

The dynamic upsampling filters (DUF)[15] (Jo et al. 2018) has been proposed, as shown in Fig. 29. It is inspired by the dynamic filter network (Jia et al. 2016) that can generate

[14] Code: https://github.com/linchuming/FFCVSR.
[15] Code: https://github.com/yhjo09/VSR-DUF.

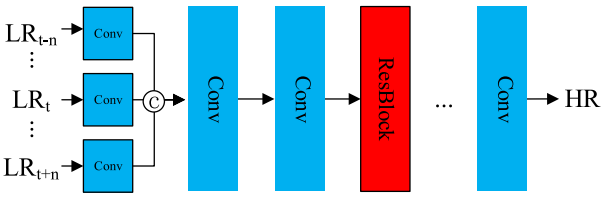

Fig. 27 The architecture of the generator in VSRResFeatGAN (Lucas et al. 2019)

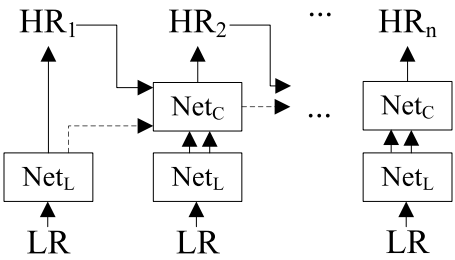

Fig. 28 The architecture of FFCVSR (Yan et al. 2019). Here Net$_C$ is the context network, and Net$_L$ is the local network

corresponding filters for specific inputs and then apply them to generate corresponding feature maps.

The structure of the dynamic up-sampling filter, together with the spatio-temporal information learned by 3D convolution, can avoid the use of motion estimation and motion compensation. DUF performs not only filtering, but also the up-sampling operation. In order to enhance high-frequency details of the super-resolution result, DUF uses a network to estimate residual map for the target frame. The final result is the sum of the residual map and the LR target frame processed by the dynamic upsample module with learned filters.

5.2.2 FSTRN

The fast spatio-temporal residual network (FSTRN) (Li et al. 2019b) uses a factorized 3D convolution to extract information contained in consecutive frames, as shown in Fig. 30. In FSTRN, a $k \times k \times k$ 3D convolutional kernel is decomposed into 2 cascaded kernels, whose sizes are $1 \times k \times k$ and $k \times 1 \times 1$, respectively, to reduce the computation caused by directly using the 3D convolution.

FSTRN consists of the following four parts: an LR video shallow feature extraction net (LFENet), fast spatio-temporal residual blocks (FRBs), an LR feature fusion and up-sampling SR net (LSRNet), and a global residual learning (GRL) module. The GRL is mainly composed of two parts: LR space residual learning (LRL) and cross-space residual learning (CRL). The LRL is introduced along with the FRBs. And the CRL directly maps the LR video to the HR space. The designs of CRL and LRL can communicate the LR and HR space. Besides, FSTRN adopts a dropout layer after LRL to enhance generalization ability of the network. LFENet using 3D convolution to extract features for consecutive LR input frames. FRBs, including the decomposed 3D convolutional layers, are responsible for extracting spatio-temporal information contained in input frames. LSRNet is used to fuse information from previous layers and conducting up-sampling.

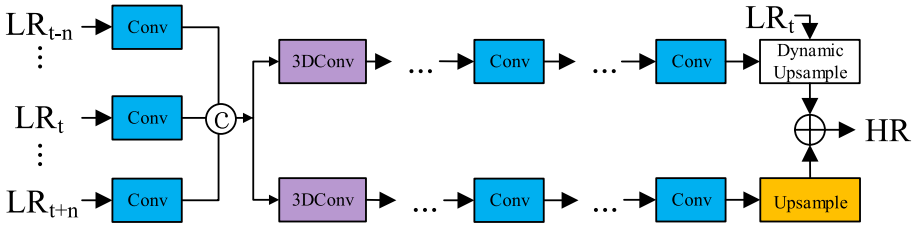

Fig. 29 The network architecture of DUF (Jo et al. 2018).

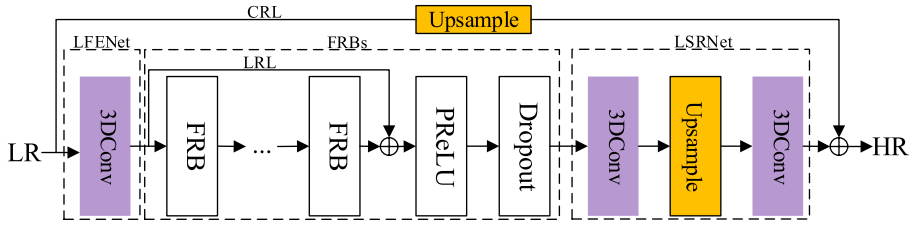

Fig. 30 The network architecture of FSTRN (Li et al. 2019b). Here FRB denotes the fast spatio-temporal residual block

5.2.3 3DSRNet

The 3D super-resolution network (3DSRNet)[16] (Kim et al. 2019) uses 3D convolution to extract spatio-temporal information contained in consecutive frames for VSR tasks. The network architecture is shown in Fig. 31. The sub net of 3DSRNet can preprocess scene change as shown in the figure. When frames of five different scenes getting involved into convolution, the sub net classifies the exact location of the scene boundary through the module of scene boundary detection, and replaces the different scene frames with the temporally closest frame of the same scene as the current middle frame. Finally, the updated five frames are sent for subsequent video super-resolution sub network. This approach overcomes performance degradation caused by scene change to some extent.

5.2.4 DSMC

A deep neural network with Dual Subnet and Multi-stage Communicated Upsampling (DSMC)[17] (Liu et al. 2021b) is proposed for super-resolution of videos with large motion. The architecture is shown in Fig. 32. It designs a U-shaped residual dense network with 3D convolution (U3D-RDN) for fine implicit MEMC as well as coarse spatial feature extraction. moreover, DSMC presents a new Multi-Stage Communicated Upsampling (MSCU) module to make full use of the intermediate results of upsampling for

[16] Code: https://github.com/sooyekim/3DSRnet.
[17] Code: https://github.com/iPrayerr/DSMC-VSR.

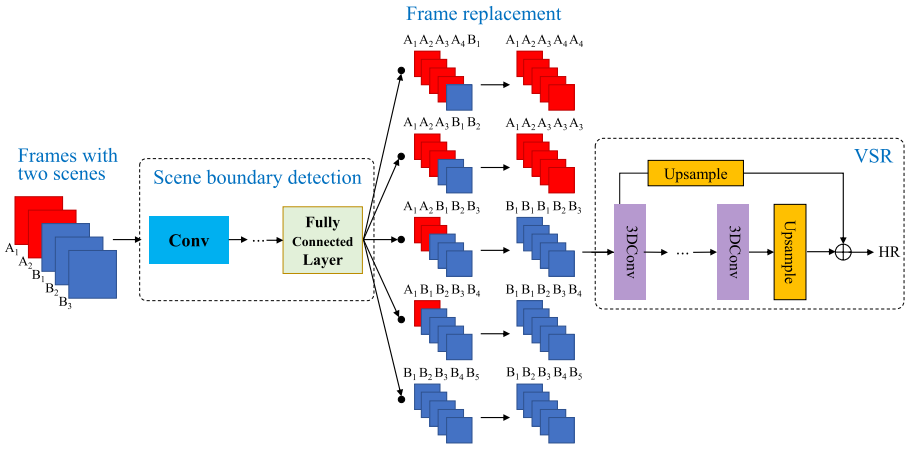

Fig. 31 The network architecture of 3DSRNet (Kim et al. 2019)

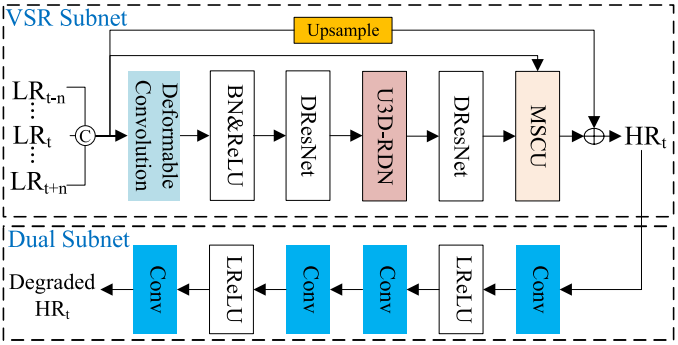

Fig. 32 The network architecture of DSMC (Liu et al. 2021b)

guiding the VSR. Besides, a dual subnet is devised to aid the training of DSMC, whose dual loss helps to reduce the solution space and enhance the generalization ability.

DSMC firstly performs deformable convolution on input consecutive frames for coarse feature extraction. The output feature maps are then processed by a deformable residual network (DResNet) (Lei and Todorovic 2018) to extract fine spatial information. Next, the feature maps are input to U3D-RDN for dimension reduction and correlation analyzation of spatio-temporal feature. Followed by another DResNet module, the feature maps are sent to MSCU module. Finally, with the aid of a dual subnet for training, DSMC yields the super-resolved HR frames. It is noted that only the output of the dual subnet and the result of VSR subnet are used for the loss computation of DSMC.

In brief, these 3D convolutional methods can extract spatio-temporal correlations contained in consecutive frames, rather than perform the motion estimation to extract motion information contained in frames and motion compensation to align them. However, most of the methods have relatively higher computational complexities compared with those of 2D convolutional methods, which limits them for real-time video super-resolution tasks.

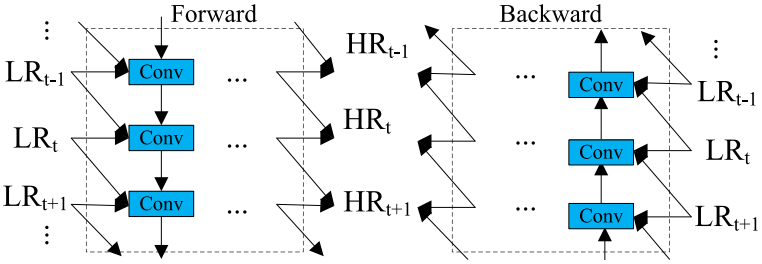

Fig. 33 The network architecture of BRCN (Huang et al. 2015, 2018)

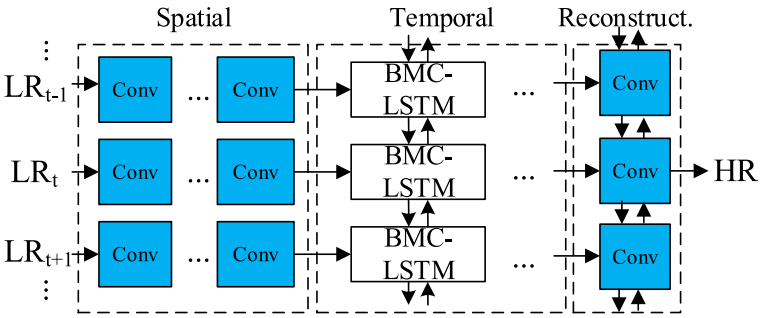

Fig. 34 The network architecture of STCN (?). Here BMC denotes the bidirectional multi-scale convolution

5.3 Recurrent convolutional neural networks (RCNNs)

It is well known that RCNNs have strong capacity in modelling temporal dependency in sequential data, e.g., natural language, video and audio. A straightforward way is to use RCNNs to handle video sequences. Based on this key idea, several RCNN methods such as BRCN (Huang et al. 2015, 2018), STCN (Guo and Chao 2017), and RISTN (Zhu et al. 2019b) have been proposed for video super-resolution.

5.3.1 BRCN

The bidirectional recurrent convolutional network (BRCN) (Huang et al. 2015, 2018), as shown in Fig. 33, is composed of two modules: a forward sub-network and a backward one with a similar structure, which only differ in the order of processing sequence. The forward subnet is responsible for modeling the temporal dependency from previous frames, while the backward subnet models temporal dependency from subsequent frames.

5.3.2 STCN

The spatio-temporal convolutional network (STCN) (Guo and Chao 2017) is an end-to-end VSR method without MEMC, as shown in Fig. 34. The temporal information within frames is extracted by using LSTM (Hochreiter and Schmidhuber 1997). Similar to RISTN (Zhu et al. 2019b), the network consists of three parts: a spatial module, a temporal module

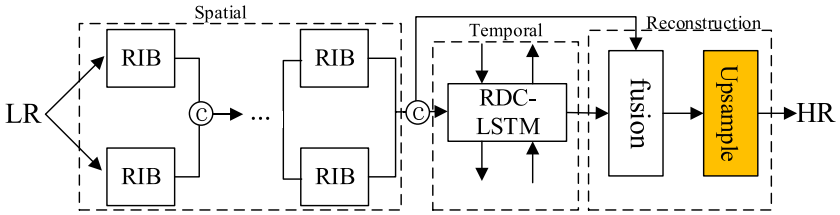

Fig. 35 The network architecture of RISTN (Zhu et al. 2019b), where RIB denotes the residual invertible block, and RDC is the residual dense convolution

and a reconstruction module. Spatial module is responsible for extracting features from multiple consecutive LR frames. Temporal module is a bidirectional multi-scale convoluted variant of LSTM, and is designed for extracting temporal correlation among frames.

5.3.3 RISTN

The residual invertible spatio-temporal network (RISTN)[18] (Zhu et al. 2019b) is inspired by the invertible block (Jacobsen et al. 2018). As shown in Fig. 35, it designs a residual invertible block (RIB), a LSTM with residual dense convolution (RDC-LSTM), and a sparse feature fusion strategy to adaptively select useful features. Here RIB is used to extract spatial information of video frames effectively, and RDC-LSTM is used to extract spatio-temporal features.

The network is mainly divided into three parts: a spatial module, a temporal module and a reconstruction module. The spatial module is mainly composed of multiple parallel RIBs, and its output is used as the input of the temporal module. In the temporal module, after extracting spatio-temporal information, features are selectively fused by a sparse fusion strategy. Finally, the HR result of the target frame is reconstructed by the deconvolution in the reconstruction module.

5.3.4 RLSP

The Recurrent Latent Space Propagation (RLSP)[19] (Fuoli et al. 2019a) shown in Fig. 36 proposes a recurrent video super-resolution algorithm, which avoids the problem that a single video frame is processed multiple times in a non-recurrent network. In addition, the algorithm implicitly transmits temporal information by introducing hidden states containing the temporal information yielded in previous moment as part of the input at current moment, and does not include explicit motion estimation and motion compensation.

The hidden state is generated by the RLSP Cell, which is composed of several convolutions. The cell receives the hidden state of the previous moment, the super-resolved result of the previous moment, as well as the current frame and the adjacent frames as inputs to yield the super-resolved result and the hidden state of the current moment. This procedure repeats until all frames are processed.

[18] Code: https://github.com/lizhuangzi/RISTN.
[19] Code: https://github.com/dariofuoli/RLSP.

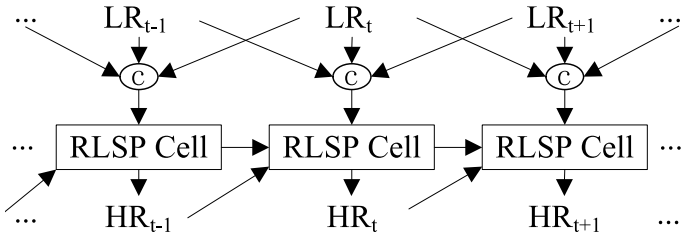

Fig. 36 The network architecture of RLSP (Fuoli et al. 2019a)

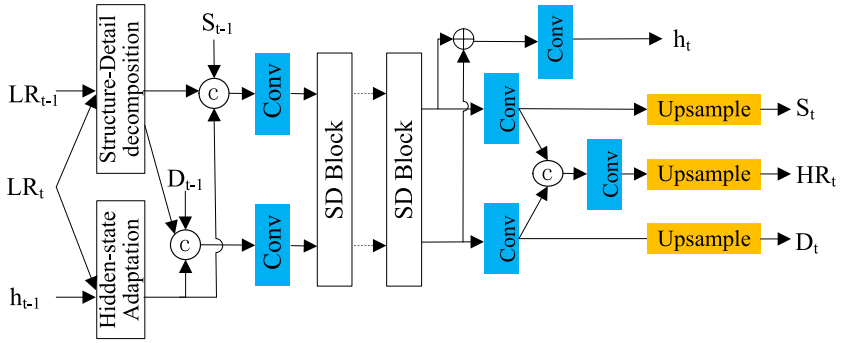

Fig. 37 The network architecture of RSDN (Isobe et al. 2020)

5.3.5 RSDN

The Recurrent Structure-Detail Network (RSDN)[20] (Isobe et al. 2020) shown in Fig. 37 proposes to divide the frame into two components, namely structure and detail, and then process these two by subsequent module respectively.

The algorithm first uses the Bicubic interpolation algorithm to downsample and upsample the input LR frame to extract the structure and detail components. Then these two components are processed by convolution and multiple SD blocks to obtain the structure and detail components, super-resolved results and hidden states at the current moment. The SD block promotes information exchange between structure and detail components. In addition, the RSDN proposes a hidden-state adaption module to select the information that is beneficial to the super resolution and avoid the interference of redundant information.

In summary, the RCNN-based methods are suitable for modeling the spatio-temporal information contained in videos, since they can map neighboring frames and thus effectively establish long-term dependence with more lightweight structures. However, conventional RCNN-based methods are difficult to train and sometimes suffer from the gradient vanishing problem. And they may not capture long-term dependence when the length of input sequences is too large, and thus may not achieve great performance. LSTM-based methods can overcome these constraints to some extent with the help of the memorization

[20] Code: https://github.com/junpan19/RSDN.

of features from shallower layers. However, the complex design of LSTM is a factor that limits their depth on hardware, restraining them to model very long-term dependence.

5.4 Non-local methods

The non-local-based method is another one that utilizes both spatial and temporal information contained in video frames for super-resolution. This method benefit from the key idea of the non-local neural network (Wang et al. 2018), which was proposed to capture long-range dependencies for video classifications. It overcomes the flaws that convolution and recurrent computations are limited to the local area. Intuitively, a non-local operation is to calculate the response value of a position, which is equal to the weight sum of all possible positions in the input feature maps. Its formula is given as follows:

$$y_i = \frac{1}{\mathcal{C}(x)} \sum_{\forall j} f(x_i, x_j) g(x_j) \tag{9}$$

where i is the index of the output location where the response value needs to be calculated, j is the index of all possible locations, x and y are the input and output data with the same dimensions, f is a function to calculate the correlation between i and j, g is the function which calculates the feature representation of input data and $\mathcal{C}(x)$ is the normalization factor. Here, g is usually defined as: $g(x_j) = W_g x_j$, where W_g is the weight matrix that needs to learn. It should be noted that f has multiple choices such as Gaussian, dot product, and concatenation. Therefore, the non-local block can easily be added into existing deep CNNs.

5.4.1 PFNL

The progressive fusion non-local (PFNL) (Yi et al. 2019) method is illustrated in Fig. 38. It mainly includes three parts: a non-local resblock, progressive fusion residual blocks (PFRB) and an upsampling block.

PFNL uses non-local residual blocks to extract spatio-temporal features, and PFRB is proposed to fuse them. Finally, the output through a sub-pixel convolutional layer is added to the input frame that is up-sampled by the bicubic interpolation, which is the final super-resolution result. PFRB is composed of three convolutional layers. Firstly, the input frames are convoluted with the 3×3 kernels, respectively, then the output feature maps are concatenated, and the channel dimension is reduced by performing the 1×1 convolution. And the results are concatenated with the previous convoluted feature maps, respectively, and conducted with a 3×3 convolution. The final results are added to each input frame to obtain the output for current PFRB.

5.5 Other

The methods in this sub-category do not utilize the initial feature extractions mentioned above. They may combine multiple techniques for super-resolution.

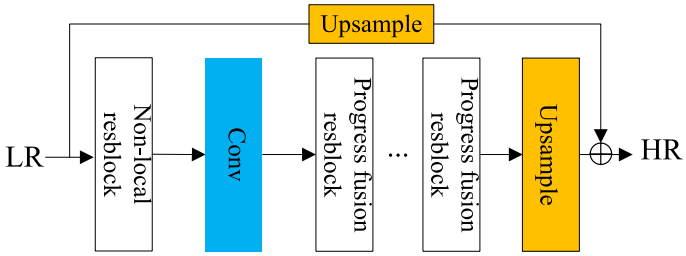

Fig. 38 The network architecture of PFNL (Yi et al. 2019)

5.5.1 RBPN

The recurrent back-projection network (RBPN)[21] (Haris et al. 2019) is inspired by the back-projection algorithm (Irani and Peleg 1991, 1993; Haris et al. 2018). RBPN mainly consists of one feature extraction module, a projection module, and a reconstruction module, and its architecture is shown in Fig. 39.

The feature extraction module includes two operations: One is to extract the features of the target frame, and the other is to extract the feature from the concatenation of the target frame, the neighboring frame, and the calculated optical flow which is from the neighboring frame to the target frame, and then perform alignment implicitly. The optical flow is obtained by the pyflow[22] method. The projection module consists of an encoder and a decoder. The encoder is composed of a multiple image super-resolution (MISR), a single image super-resolution (SISR) and residual blocks (denoted as ResBlock). The decoder consists of ResBlock and a strided convolution, and it takes the output of the previous encoder as input to produce LR features for the encoder of the next projection module. The concatenation of the target frame, the next neighboring frame and pre-computed optical flow are input to the feature extraction module, whose output is also for the encoder in the next projection module. The above process does not stop until all neighboring frames are processed. That is, projection is used recurrently, which is the reason of the words "recurrent back-projection network". Finally, the reconstruction module takes the output of the encoder in each projection module by the mean of concatenation as input to produce the final SR result.

5.5.2 STARnet

The architecture of space-time-aware multi-resolution networks (STARnet) (Haris et al. 2020) is shown in Fig. 40. STARnet is an end-to-end network that can simultaneously process video super-resolution and video interpolation. It consists of the following three stages: initialization, refinement and reconstruction.

In the initialization stage, STARnet receives four parts of inputs including two LR RGB frames and their bidirectional flow images. In this stage, the two spatial super-resolution (S-SR) modules can execute super-resolution to the two LR frames by DBPN (Haris et al.

[21] Code: https://github.com/alterzero/RBPN-PyTorch.
[22] https://github.com/pathak22/pyflow.

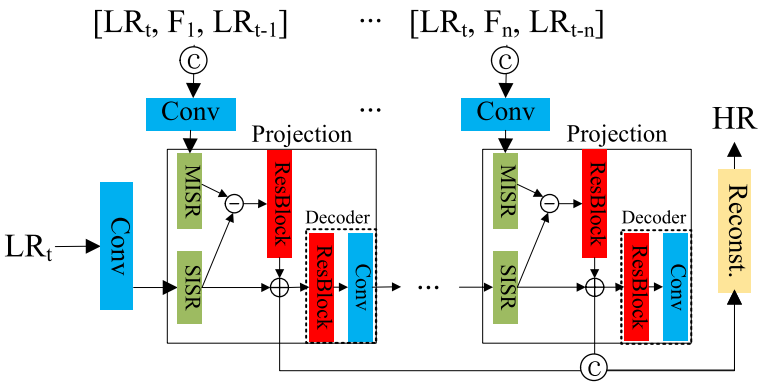

Fig. 39 The network architecture of RBPN (Haris et al. 2019), where © denotes concatenation, ⊖ is element subtraction, and MISR denotes multi-image super-resolution

2018) or RBPN (Haris et al. 2019) and re-generate their LR counterparts by a similar network to prepare for frame interpolation in both LR and HR spaces in the spatio-temporal super-resolution (ST-SR) module. Meanwhile, the motion module align the bidirectional flow images.

5.5.3 DNSTNet

DNSTNet (Sun et al. 2020) is the video super-resolution via the Dense Non-local Spatial-Temporal convolutional Network, shown in Fig. 41. Here, the dense feature extraction subnetwork composes of the Short-term Temporal Dependency Extraction Block (S-TBlock), Long-term TBlock (L-TBlock), and dense connections, as shown in this figure. It utilizes 3D convolution to capture short-term temporal dependency existing from adjacent frames in S-TBlock, and the bidirectional ConvLSTM for capturing long-term temporal dependency in L-TBlock. It also proposes a region-level non-local block following the dense feature extraction to exploit the global information, and to enlarge the limited receptive field of 3D convolution and ConvLSTM. This non-local network divides the feature maps into multiple patches and processes them respectively to decrease the computational cost. To sum up, DNSTNet adopts multiple modules to improve the performance of VSR.

Although DNSTNET uses a 3D convolution module, an LSTM module and a non-local sub-network, it does not imply better performance than EDVR and DSMC. As it is known, the network architecture is one of the important factors to affect its performance, other factors including the training strategy, and iteration number also influence its performance. Compared with the methods: EDVR and DSMC, the training strategy of DNSTNET is probably not sophisticated designed. It is a common initializing method. But EDVR is initialized by parameters from a shallower similar network. This can boost the performance. DSMC also has a deeper structure, which may be helpful to improve the performance. Moreover, in DNSTNET, too many features without selection through the dense feature concatenation are input to the non-local block for computation. These features may bring redundant information, which results in performance degradation. While in DSMC, the extracted features are refined through the U3D-RDN module before they are input to a non-local block. This processing can enhance the performance.

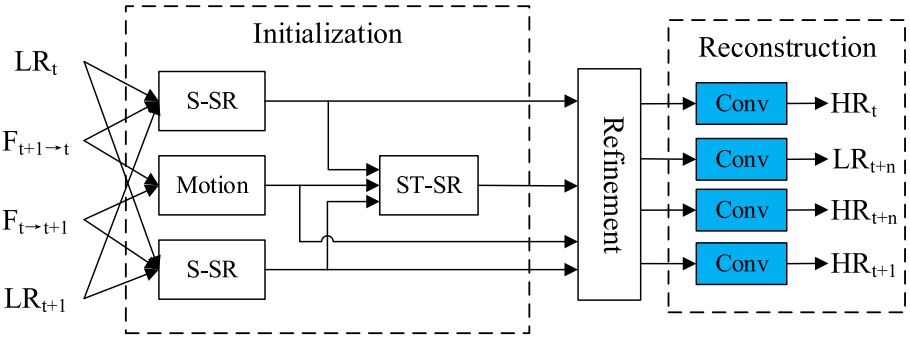

Fig. 40 The network architecture of STARnet (Haris et al. 2020)

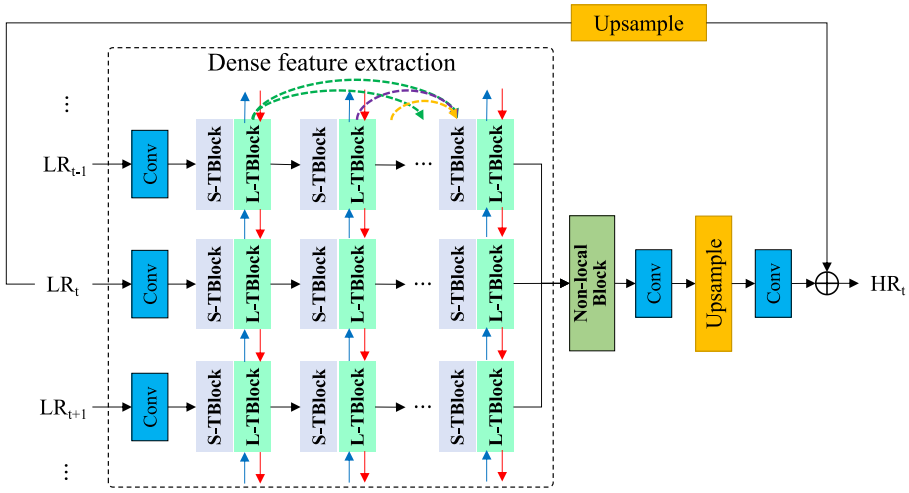

Fig. 41 The network architecture of DNSTNet (Sun et al. 2020)

In summary, the non-local based methods introduce attention mechanisms into VSR tasks. They can establish effective dependence of spatio-temporal information by extending the receptive field to the global. However, the non-local modules used in them need to calculate the response at each position by attending to all other positions and computing a weighted average of the features in all positions. Thus, this incurs high computational cost, and some efforts can be made to reduce the computational overhead of the methods.

Moreover, the methods without alignment rely on the non-linear capability of the neural network to learn the motion correlation between frames for video super-resolution. They do not utilize additional modules to align frames. The learning ability largely depends on the design of the deep neural network. And an elaborate design is more likely leading to higher performance for video super-resolution.

In addition, we discuss the deeper connections between all the methods below. (1) The methods such as EDVR, DNLN, TDAN, D3Dnet and VESR-Net, which belong to the deformable convolution category, all attempt to overcome the flaw of optical flow-based methods by using the DConv structure. The estimation of optical flow is inaccurate when

Table 2 Some widely used video super-resolution datasets

Dataset	Year	Type	Download link	Video number	Resolution	Color space
YUV25	*	Train	https://media.xiph.org/video/derf/	25	386 × 288	YUV
TDTFF	*	Test	www.wisdom.weizmann.ac.il/~vision/SingleVideoSR.html	5	648 × 528 for Turbine, 960 × 530 for Dancing, 700 × 600 for Treadmill, and 1000 × 580 for Flag, 990 × 740 for Fan	YUV
Vid4	2011	Test	https://drive.google.com/drive/folders/10-gUO6zBeOpWEamrWKCtSkkUFukB9W5m	4	720 × 480 for Foliage and Walk, 720 × 576 for Calendar, and 704 × 576 for City	RGB
YUV21	2014	Test	http://www.codersvoice.com/a/webbase/video/08/152014/130.html	21	352 × 288	YUV
Venice	2014	Train	https://www.harmonicinc.com/free-4k-demo-footage/	1	3840 × 2160	RGB
Myanmar	2014	Train	https://www.harmonicinc.com/free-4k-demo-footage/	1	3840 × 2160	RGB
CDVL	2016	Train	http://www.cdvl.org/	100	1920 × 1080	RGB
UVGD[a]	2017	Test	http://ultravideo.cs.tut.fi/	7	3840 × 2160	YUV
LMT[b]	2017	Train	http://mcl.usc.edu/mcl-v-database, http://live.ece.utexas.edu/research/quality/live_video.html,https://vision.in.tum.de/data/datasets	*	1920 × 1080	RGB
Vimeo-90K	2019	Train+Test	http://toflow.csail.mit.edu/	91,701	448 × 256	RGB
REDS[c]	2019	Train+Test	https://seungjunnah.github.io/Datasets/reds.html	270	1280 × 720	RGB

Note that '*' represents unknown information

[a] UVGD denotes the Ultra Video Group Database

[b] LMT denotes the LIVE video quality assessment database, the MCL-V database, and the TUM 1080p dataset.

[c] REDS denotes the REalistic and Diverse Scenes

Table 3 Some major video super-resolution competitions

Name	Year	Organizer	Location	Website	Dataset	Scale	Champion	PSNR	SSIM
NTIRE 2019 Video Restoration and Enhancement Challenges	2019	CVPR	Long Beach, California	https://data.vision.ee.ethz.ch/cvl/ntire19/	REDS	×4	EDVR (Wang et al. 2019c)	31.79	0.8962
YOUKU Video Super-Resolution and Enhancement Challenge	2019	Alibaba	Hangzhou, China	https://tianchi.aliyun.com/competition/entrance/231711/introduction	Youku-VESR	×4	VESR-Net (Chen et al. 2020)	37.85	*
AIM 2019 Challenge on Video Extreme Super-Resolution	2019	ECCV	Hong Kong, China	https://www.aim2019.org/	Vid3oC	×16	EDVR+	22.53	0.6400
Mobile Video Restoration Challenge	2019	ICIP & Kwai	*	https://www.kuaishou.com/activity/icip2019	*	*	*	*	*
AIM 2020 Challenge on Video Extreme Super-Resolution	2020	ECCV	Boston, Massachusetts	http://aim2020.org/	Vid3oC	×16	EVESRNet (Dario et al. 2020)	22.83	0.6450
Mobile AI 2021 Real-Time Video Super-Resolution Challenge	2021	CVPR	VIRTUAL	https://ai-benchmark.com/workshops/mai/2021/	REDS	×4	Diggers (Ignatov et al. 2021)	28.33	0.8112
NTIRE 2021 Video Super-Resolution Challenge	2021	CVPR	VIRTUAL	https://data.vision.ee.ethz.ch/cvl/ntire21/	REDS	×4	BasicVSR++ (Chan et al. 2021c)	33.36	0.9218

Note that 'EDVR+' stands for a method based on EDVR, and '*' represents unknown information

dealing with complex motions and varying illumination, while the receptive field of the convolution kernel can be expanded by utilizing DConv. And the network can better capture complex motions and illumination changes. (2) The methods such as DUF, FSTRN, 3DSRnet and DSMC all employ 3D convolutional layers to learn spatial and temporal features simultaneously instead of the 2D convolution from video data. Besides, they also try to avoid the inaccuracy of motion estimation and compensation when complex motions involve by designing new network structures. (3) The methods such as BRCN, STCN, RISTN, RLSP, RSDN and BasicVSR exploit long-term contextual information contained in video frames by using bidirectional recurrent convolutional networks. The bidirectional RCNN can utilize temporal dependency from both previous and future frames through the combination of a forward recurrent network and a backward recurrent network. (4) The methods such as RVSR, STCN, BRCN, EDVR, DNLN, TDAN, D3DNet, VESR-Net, DUF, 3DSRNet and DSMC involve the dealing with complex motions in the videos. (5) The methods such as MuCAN (in MEMC class), EDVR (in DC class), VESR-Net (in DC class) and PFNL (in non-local class) attempt to capture the global dependency between different positions across frame. Specifically, the TSA module in the EDVR method assigns pixel-level weights on each frame for fusion. MuCAN, VESR-Net, and PFNL all design non-local modules to correlate different patches, which improve the ability to capture motion information. (6) The methods such as DRVSR, MultiBoot VSR, and DSMC all address the video super-resolution with multiple scaling factors. They not only consider ×4 scale, but also regard ×2, ×3, or ×8 scales. (7) The methods such as MultiBoot VSR, PFNL and RBPN all pay attention to improve the training strategies. For example, PFNL adopts residual learning to stabilize the training process.

6 Performance comparisons

6.1 Datasets and competitions

Details of some of the most popular datasets used in VSR tasks are summarized in Table 2. The most widely-used dataset for training is Vimeo-90K, since it is currently the largest VSR dataset with real scenes. The most popular dataset for testing is Vid4, whose frames contain more high-frequency details than others. Thus, Vid4 is frequently used for evaluating the performance of VSR methods. REDS includes videos with extremely large movement, which is challenging for VSR methods.

Besides, we also summarize several international competitions on video super-resolution in Table 3. The NTIRE 2019 Challenge (Nah et al. 2019a, b) aims at recovering videos with large movements and diverse real-world scenes. Its winning solution is EDVR (Wang et al. 2019c), which may be one of the most popular works for VSR. The AIM Challenges in 2019 (Fuoli et al. 2019b) and 2020 (Fuoli et al. 2020) both encourage solutions of VSR with large scale factors. A method enhanced from EDVR won the AIM 2019 Challenge, while EVESRNet (Dario et al. 2020) won the AIM 2020 Challenge. Besides, the YOUKU Video Super-Resolution and Enhancement Challenge, and Mobile Video Restoration Challenge in 2019 are both for videos which are more relevant to entertainment. The winning solution of YOUKU challenge is VESR-Net (Chen et al. 2020). The Mobile AI 2021 Real-Time Video Super-Resolution challenge (Ignatov et al. 2021) evaluated the solutions on an OPPO Find X2 smartphone GPU. The most recent NTIRE 2021 Challenge

Table 4 Comparison of all the methods on the datasets with scale factor ×4

Method	Training Set	Test Set	Channel	Params.(Million)	BI		BD	
					PSNR	SSIM	PSNR	SSIM
Deep-DE	*	city+temple+penguin	*	1.11[1]	–	–	29.00	0.8870
VSRnet	Myanmar	Vid4	Y	0.27[2]	24.84	0.7049	–	–
		Myanmar–T	Y		31.85	0.8834	–	–
VESPCN	CDVL	Vid4	Y	0.88[2]	25.35	0.7557	–	–
DRVSR	*	Vid4	Y	2.17[3]	25.52	0.7600	–	–
		SPMCS	Y		29.69	0.8400	–	–
RVSR	LMT	Vid4+temple+penguin	Y	–	28.05	–	–	–
		UVGD	Y		39.71	–	–	–
FRVSR	Vimeo-90K	Vid4	Y	2.81[3]	–	–	26.69	0.8103
		Vimeo-90K-T	Y		–	–	35.64	0.9319
SOFVSR	CDVL	DAVIS-10	Y	1.71[3]	34.32	0.9250	34.27	0.9250
		Vid4	Y		26.01	0.7710	26.19	0.7850
TecoGAN	*	ToS	Y	3.00	–	–	32.75	–
		Vid4	Y		–	–	25.89	–
TOFlow	Vimeo-90K	Vid4	Y	1.41	23.54	0.8070	–	–
		Vimeo-90K-T	Y		33.08	0.9417	–	–
MMCNN	*	Vid4	Y	10.58	26.28	0.7844	–	–
		Myanmar-T	Y		33.06	0.9040	–	–
		YUV21	Y		28.90	0.7983	–	–
		Vid4+temple+penguin	Y		28.97	–	–	–
MEMC-Net	Vimeo-90K	Vimeo-90K-T	Y	–	33.47	0.9470	–	–
		Vid4	Y		24.37	0.8380	–	–
RRCN	Myanmar	Myanmar-T	Y	–	32.35	0.9023	–	–
		Vid4	Y		25.86	0.7591	–	–
		YUV21	Y		29.08	0.7986	–	–
RTVSR	harmonicinc.com	Vid4	Y	15.00	26.36	0.7900	–	–
		Vid4+temple+penguin	Y		29.03	–	–	–

Table 4 (continued)

Method	Training Set	Test Set	Channel	Params.(Million)	BI		BD	
					PSNR	SSIM	PSNR	SSIM
MultiBoot VSR	REDS	REDS-T	RGB	60.86	31.00	0.8822	–	–
MuCAN	Vimeo-90K	Vimeo-90K-T	Y	19.90	37.32	0.9465	–	–
	REDS	REDS4	RGB	25.70	30.88	0.8750	–	–
IconVSR	Vimeo-90K	Vimeo-90K-T	Y	8.70	37.47	0.9476	37.84	0.9524
		Vid4	Y		27.39	0.8279	28.04	0.8570
		UDM10	Y		–	–	40.03	0.9694
	REDS	REDS4	RGB		31.67	0.8948	–	–
EDVR	Vimeo-90K	Vid4	Y	20.60	27.35	0.8264	27.85	0.8503
		Vimeo-90K-T	Y		37.61	0.9489	37.81	0.9523
	REDS	REDS4	RGB		31.09	0.8800	28.88	0.8361
DNLN	Vimeo-90K	Vid4	Y	19.74	27.31	0.8257	–	–
		SPMCS	Y		30.36	0.8794	–	–
TDAN	Vimeo-90K	Vid4	Y	1.97[2]	26.24	0.7800	26.58	0.8010
D3Dnet	Vimeo-90K	Vid4	Y	2.58	26.52	0.7990	–	–
VESR-Net	Youku-VESR	Youku-VESR-T	RGB	21.65	–	–	35.97	–
VSRResFeatGAN	Myanmar	Vid4	Y	–	25.51	0.7530	–	–
FFCVSR	Venice+Myanmar	Vid4	Y	–	26.97	0.8300	–	–
DUF	Vimeo-90K	Vid4	Y	5.82	–	–	27.38	0.8329
		Vimeo-90K-T	Y		–	–	36.87	0.9447
	REDS	REDS4	Y		28.63	0.8251	–	–
FSTRN	YUV25	TDTFF	Y	–	–	–	29.95	0.8700
3DSRnet	largeSet	Vid4	Y	0.11[3]	25.71	0.7588	–	–
DSMC	REDS	REDS4	RGB	11.58	30.29	0.8381	–	–
		Vid4	Y		27.29	0.8403	–	–

Table 4 (continued)

Method	Training Set	Test Set	Channel	Params.(Million)	BI		BD	
					PSNR	SSIM	PSNR	SSIM
BRCN	YUV25	Vid4	Y	–	–	–	24.43	0.6334
		TDTFF	Y		–	–	28.20	0.7739
STCN	*	Hollywood2	Y	–	–	–	34.58	0.9259
		city+temple+penguin	*		–	–	30.27	0.9103
RISTN	Vimeo-90K	Vid4	Y	3.67	26.13	0.7920	–	–
RLSP	Vimeo-90K	Vid4	Y	4.21	–	–	27.48	0.8388
		Vimeo-90K-T	Y		–	–	36.49	0.9403
RSDN	Vimeo-90K	Vid4	Y	6.19	–	–	27.92	0.8505
		Vimeo-90K-T	Y		–	–	37.23	0.9471
		UDM10	Y		–	–	39.35	0.9653
PFNL	Vimeo-90K	Vid4	Y	3.00	26.73	0.8029	27.16	0.8355
		Vimeo-90K-T	Y		36.14	0.9363	–	–
	REDS	REDS4	RGB		29.63	0.8502	–	–
RBPN	Vimeo-90K	Vid4	Y	12.20	27.12	0.8180	–	–
		Vimeo-90K-T	Y		37.07	0.9453	37.20	0.9458
	REDS	REDS4	RGB		30.09	0.8590	–	–
STARnet	Vimeo-90K	UCF101	*	111.61[4]	29.11	0.9240	–	–
		Vimeo-90K-T	*		30.83	0.9290	–	–
		Middlebury	*		27.16	0.8270	–	–
DNSTNet	Vimeo-90K	Vid4	Y	–	27.21	0.8220	–	–
		Vimeo-90K-T	Y		36.86	0.9387	–	–
		SPMCS	Y		29.74	0.8710	–	–

Note that 'Internet' means that the dataset is collected from the internet. '*' denotes that the source of the dataset is unknown, and '–' indicates that the method does not be tested on the datasets. Here, the numbers of the parameters in the models marked with [1], [2], [3] and [4] are from Wang et al. (2019b), Ying et al. (2020), Liu et al. (2021b), and Isobe et al. (2020), respectively

on Video Super-Resolution gauges the state-of-the-art (Son et al. 2021), its winner being BasicVSR++ (Chan et al. 2021c). These competitions are making great contributions to the development of video super-resolution and helping develop new methods for various video super-resolution applications.

6.2 Performance of methods

Moreover, we summarize the performance of the representative VSR methods with scale factor 4 in Table 4 in terms of both PSNR and SSIM. More experimental results for VSR tasks with magnification factors 2 and 3 are reported in Supplementary Materials. The degradation types are the bicubic downsampling with the image-resize function (BI) and Gaussian blurring and downsampling (BD). Note that part of the PSNR and SSIM are from their original works. And a simple comparison on the performance may not be fair, since the training data, the pre-processing, and the cropped area in videos are likely totally different in the methods. The details about the performance are listed to provide reference for readers.

According to Table 4, the top 5 methods in the ×4 VSR task on Vimeo-90K-T dataset are as follows. The methods are denoted by (PSNR, BI/BD, Params.). IconVSR (37.84, BD, 8.70), EDVR (37.61, BI, 20.60), IconVSR (37.47, BI, 8.70), MuCAN (37.32, BI, 19.90), and RSDN (37.23, BD, 6.19). The top 5 methods on Vid4 dataset are IconVSR (28.04, BD, 8.70), RSDN (27.92, BD, 6.19), EDVR (27.85, BD, 20.60), RLSP (27.48, BD, 4.21), and DUF (27.38, BD, 5.82). The top 4 methods on REDS4 dataset are IconVSR (31.67, BI, 8.70), EDVR (31.09, BI, 20.60), MuCAN (30.88, BI, 25.70), and DSMC (30.29, BI, 11.58). In the method evaluation, we compare the results on Y channel for Vimeo-90K-T and Vid4 datasets, and on RGB channel for REDS4. PFNL and DNLN do not utilize all the test frames.

IconVSR and EDVR show superior performance on the three datasets. And IconVSR uses optical flow for feature alignment, a bidirectional recurrent network for temporal feature propagation, and an information-refill mechanism for feature refinement. With these properties, it outperforms some other methods in some cases, and achieves more performance gain with BD degradation than BI degradation on Vimeo-90K-T and Vid4. EDVR employs cascaded multi-scale deformable convolutions for alignment, and TSA to fuse multiple frames. Unlike DNLN, which also adopts deformable convolutions, EDVR can capture multi-scale feature information. Compared with TDAN and D3Dnet, the architecture of EDVR is more complicated and may learn more information from inputs, though they all employ deformable convolutions for alignment. And EDVR costs 20.60 million (M) parameters, which is far more than other top networks. This may explain its better performance.

For the Vid4 dataset, RLSP and RSDN both adopt recurrent convolutional neural network as backbone to utilize the temporal information contained in multiple frames. RSDN further divides a frame into structure and detail to process them respectively, and also exchange the information between them. This refined extraction attributes to its performance. PFNL proposes the non-local residual block to capture long-range spatio-temporal dependencies between frames, which may outperform some conventional MEMC-based methods.

For the Vimeo-90K-T dataset, the performance of MuCAN is likely attributed to the two main modules, CN-CAM and TM-CAM. The former module can hierarchically aggregate information for handling large and subtle motion, and the latter one

captures non-local communication within different feature resolutions. RSDN relies the information exchange between the structure and detail to gain better performance on Vimeo-90K-T. It is noticed that MuCAN has 19.90 M parameters, which is far more than those of RSDN and IconVSR on this dataset.

Moreover, for the REDS4 dataset, it is noticed that EDVR and MuCAN both have more than 20.0 M parameters, which is far more than those of IconVSR and DSMC, though they are in the second and the third places on the top list. DSMC proposes the U3D-RDN module which learns coarse-to-fine spatio-temporal features, and MSCU, which decomposes an upsampling into multiple sub-tasks for full use of the intermediate results as well as a dual subnet for aiding the training. DSMC demonstrates superior performance to other 3D convolution methods.

6.3 Guidelines for model selection

In this subsection, we provide some guidelines for readers to select different models according to the results in Table 4. For the super-resolving videos with realistic textures and rich details but without large motions, the following methods can be prime candidates: IconVSR, RSDN, EDVR, RLSP, DUF, DNLN, DSMC, PFNL, RBPN, and FRVSR. These methods are ordered according to the PSNR values on the Vid4 dataset, whose videos contain more high-frequency details. Among them, EDVR and DNLN both have more than 20.0 M parameters, which are suitable for the applications without tight restriction on GPU memory. And the methods such as IconVSR, RSDN, RLSP, DUF, PFNL and FRVSR cost less than 10.0 M model parameters, which might be more appropriate for the application of mobile devices and embedding systems.

When dealing with video sequences with complex and large motions, one can select the following methods, IconVSR, EDVR, DSMC, RBPN, and PFNL. The performance of these methods is ranged in descending order and referred to their PSNR results on the REDS dataset. Similar to the above applications, the number of the parameters in EDVR exceeds 20.0 MB, while those of IconVSR and PFNL are fewer than 10.0 MB.

For generic videos except for the above two videos, we recommend the methods, IconVSR, EDVR, MuCAN, RSDN, RBPN, RLSP, PFNL and FRVSR. These methods are ordered according to the PSNR values on the Vimeo-90k-T dataset. The number of the parameters in EDVR is larger than 20.0 MB, and those of IconVSR, MuCAN, RSDN, RLSP, PFNL and FRVSR are fewer than 10.0 MB.

There are some additional tips for selecting the methods with alignment. When inaccurate motion estimation and alignment may introduce artifacts for videos with large motions or lighting changes, the deformable convolution-based methods are more robust for VSR tasks. When considering the online applications of video super-resolution, a unidirectional network may be the best candidate, where the information is sequentially propagated from the first frame to the last frame. While for offline applications, a bidirectional network in which the features can propagate forward and backward in time independently, is a better choice for VSR. In this case, the optical flow can be estimated both sequentially and reversely. It is known that the motion estimation is one critical step for the methods with alignment, which directly influences the performance of VSR methods. When more advanced estimation methods are proposed, they can be used to improve VSR's performance.

7 Applications of video super-resolution

By using VSR techniques, the resolution of video frames can be enhanced, and better visual quality and recognition accuracy can be achieved. It has a variety of applications, such as remote sensing, medical diagnoses, video decoding, and 3D reconstruction.

7.1 Video decoding

In Glaister et al. (2011), a patch-based super-resolution method was presented to decode frames for video playback, and had been integrated in a video compression pipeline. Dai et al. (2015) proposed a VSR algorithm based on dictionary learning and sub-pixel motion compensation. This algorithm adopted multiple bilevel dictionaries for single-frame SR. Meanwhile, they presented a dictionary learning algorithm, where the dictionaries are trained from consecutive video frames. In Liu and Cui (2018), an improved super-resolution reconstruction algorithm, which was a part of the proposed low bit-rate coding scheme, was applied to the decoded data for reconstructing high-definition videos. In Umeda et al. (2018), an anchored neighborhood regression SR method (Timofte et al. 2014) was used for decoding in the proposed video coding system.

Kim et al. (2018b) proposed a hardware-friendly VSR algorithm which can upscale full-high-definition (FHD) video streams to their 4K ultra-high-definition counterparts, and implemented it in both field programmable gate array (FPGA) and application specific integrated circuit (ASIC) hardware for real-time video reconstruction. They further presented a FPGA-based network structure for SR. The number of parameters is reduced by using cascaded convolutions and depth-wise separable residual network (Kim et al. 2018c). In Wei et al. (2019), a CNN-based SR algorithm was implemented and accelerated through network pruning and quantization, and the algorithm was integrated in their real-time FPGA-based system, which supports video stream transcoding from H.264 FHD to H.265/HEVC UHD.

7.2 Remote sensing

The image SR methods such as VDSR and ESPCN have been utilized to enhance the resolution of objects in satellite videos in Luo et al. (2017); Xiao et al. (2018). In Jiang et al. (2018a), a progressively enhanced network with a transition unit was proposed to strengthen residual images with fine details. Moreover, Jiang et al. (2018b) proposed a deep distillation recursive network with a multi-scale purification unit to super-resolve the images in the Jilin-1 satellite videos. Liu et al. (2020a) proposed a framework to pose the image priors in maximum a posteriori to regularize the solution space and generate the corresponding high-resolution video frames. The framework combines the implicitly captured local motion information through exploiting spatio-temporal neighbors and the non-local spatial similarity to recover HR frames. The experiments on the videos from the Jilin-1 satellite and the OVS-1A satellite verify that the approach can preserve edges and texture details.

7.3 Medical analysis

Poot et al. (2010) and Odille et al. (2015) reconstructed isotropic 3D magnetic resonance imaging (MRI) data in high resolution from multiple low-resolution MRI slices of different orientations, and they did not utilize accurate motion estimation and alignment. In Zhang

et al. (2012), HR 4D computerized tomography (CT) images are super-resolved with several frames for each slice at different respiratory phases. Yu et al. (2017) proposed a multi-slice CT SR network, which inputs the consecutive CT slices as video frames. It is composed of several convolution layers and a rearranging layer, and a subset of 5800 slices is used to train the model, and the other 1,000 slices for testing.

Ren et al. (2019) proposed a framework, which adopts an iterative upsampling layer and one downsampling layer in DBPN (Haris et al. 2018) to provide an error feedback mechanism for the reconstruction of medical videos. Lin et al. (2020) proposed a network to super-resolve the cardiac MRI slices, which uses bidirectional ConvLSTM as the network backbone. It utilizes domain knowledge of cardiac, and iteratively to enhance low-resolution MRI slices.

7.4 Surveillance videos

Shamsolmoali et al. (2019) proposed a deep CNN to upsample the low-resolution surveillance videos. The CNN is composed of less than 20 layers and is trained and tested on two surveillance datasets, which are mainly indoor videos. Lee et al. (2018) utilized SRGAN (Ledig et al. 2017) to enhance the details of the characters on license plate, and they also collected a video dataset with low-resolution and evaluated their method to verify its effectiveness. Guo et al. (2020) adopted DeblurGAN (Kupyn et al. 2018) to remove the motion blur of the adjacent frames, and then the MEMC was performed on adjacent frames. Finally, high-resolution video frames can be reconstructed through a multi-frame super-resolution algorithm.

In order to super-resolve multi-view face video, Deshmukh and Rani (2019) proposed a fractional-Grey Wolf optimizer-based kernel for the neighboring pixel estimation in the face video. Xin et al. (2020) proposed a simple but effective motion-adaptive feedback cell that can capture the motion information and feed it back to the network in an adaptive way for video face super-resolution.

7.5 3D reconstruction

By using the input video sequences, Burns et al. (2017) presented a SR method, which produces a 3D mesh of the observed scene with enhanced texture. For multiview video SR methods, Li et al. (2016) adopted the kernel regression to upgrade the information extraction layer, and utilize non-local means to information merging layer. Furthermore, Li et al. (2019c) proposed the first framework that super-resolves the appearance of 3D objects that are captured from multiple view points. The framework combines the power of 2D deep learning-based techniques with the 3D geometric information in the multi-view setting.

7.6 Virtual reality

Liu et al. (2020b) proposed a single frame and multi-frame joint super-resolution network, which includes a loss function with weighted mean squared error for the SR of 360-degree panorama videos. They also provided a new panorama video dataset: the MiG Panorama Video for evaluating the panorama VSR algorithms. Dasari et al. (2020) presented a video streaming system to reduce bandwidth requirements for 360-degree videos. The client runs a deep learning based SR model to recover the video, which is heavily compressed at the

server. The authors also compared with other state-of-the-art video streaming systems on video quality of experience.

7.7 Thermal videos

In Kwasniewska et al. (2019), a super-resolution model based on CNN and residual connection was proposed to enhance the thermal videos acquired by thermal cameras, and contactlessly estimate the respiratory rate. Compared with the previous methods, the performance is improved by using super-resolved sequences. Gautam and Singh (2020) discussed the performance of SR techniques by using different deep neural networks on benchmark thermal datasets, including SRCNN (Dong et al. 2014), EDSR (Lim et al. 2017), autoencoder and SRGAN (Ledig et al. 2017). Based on the experimental results, they concluded that SRGAN yields superior performance on thermal frames when comparing with others.

8 Trends and challenges

Although great progress has been made by state-of-the-art video super-resolution methods based on deep learning especially on some public benchmark datasets, there are still challenges and trends discussed below.

8.1 Lightweight super-resolution models

The deep learning based video super-resolution methods enjoy high performance, nevertheless they have difficulty in deploying efficiently in many real-world problems. It is noted that their models usually have a mass of parameters and require vast computational and storage resources, and their training also takes a long time. With the popularity of mobile devices in modern life, one expects to apply these models on such devices. To address this issue, several lightweight super-resolution methods have been proposed, e.g., RISTN (Zhu et al. 2019b), TDAN (Tian et al. 2020), and Xiao et al. (2021). How to design and implement a lightweight super-resolution algorithm with high performance for real-world applicants is a major challenge.

8.2 Interpretability of models

Deep neural networks are usually considered as black boxes. That is, we do not know what real information the model learns when the performance is good or bad. In existing video super-resolution models, there is not a theoretical interpretation about how convolution neural networks recover low-resolution video sequences. With a deeper investigation on its interpretation, the performance of super-resolution algorithms for both videos and images may be improved greatly. Some works have paid attention to this problem, e.g., Chan et al. (2021b) and Liu et al. (2021a).

8.3 Super-resolution with larger scaling factors

For video super-resolution tasks, existing works mainly focus on the case of the magnification factors ×2, ×3 and ×4. The more challenging scales such as ×8 and ×16 have

been rarely explored. With the popularity of high-resolution (e.g., 8K and 16K) display devices, larger scaling factors are to be further studied. Obviously, as the scale becomes larger, it is more challenging to predict and restore unknown information in video sequences. This may result in performance degradation for the algorithms, and weaken robustness in the models. Therefore, how to develop stable deep learning algorithms for VSR tasks with larger scaling factors is still challenging. Until now, there is seldom such work on VSR, while several works such as Chan et al. (2021a) and Chen et al. (2021) were proposed for the single image super-resolution tank with larger scaling factors, e.g., ×8.

8.4 Super-resolution with arbitrary scaling factors

From Table 4, we can see that most video super-resolution methods are designed for the case of the scaling factor ×4, which is not appropriate for real scenes. On the one hand, other scales like ×2, ×3 or ×1.5 are also very common in VSR tasks. On the other hand, a video super-resolution model with fixed scale will seriously limit its generalization and portability. Therefore, a universal VSR method for arbitrary scale factors is greatly needed in real-world applications. Several works about image super-resolution with arbitrary scaling factors have been presented, e.g., Hu et al. (2019), and Wang et al. (2021b), while the works on arbitrary scale factor upsampling for videos are still seldom.

8.5 More reasonable and proper degradation process of videos

In existing works, the degraded LR videos are attained through the two methods: One is directly downsampling HR videos by using interpolation, such as bicubic. The other is performing the Gaussian blurring on HR videos and then downsampling the video sequences. Although both methods perform well in theory, they always perform poorly in practice. As it is known, the real-world degradation process is very complex and includes much uncertainty. The blurring and interpolation are not adequate for modeling this problem. Therefore, when constructing LR videos, the degradation should be modeled theoretically in consistent with the real-world case to reduce the gap between research and practice. There are a few works involving the degradation process of videos for super-resolution, such as Zhang et al. (2018a).

8.6 Unsupervised super-resolution methods

Most state-of-the-art VSR methods adopt a supervised learning paradigm. In other words, the deep neural networks require a large number of paired LR and HR video frames for training. However, such paired datasets are hard or costly to obtain in practice. One may synthesize the LR/HR video frames, the performance of super-resolution methods is still not satisfied as the degradation model is too simple to characterize the real-world problem and results in inaccurate HR/LR datasets. Thus, unsupervised VSR methods are highly demanded. Some works of unsupervised VSR on satellite videos have been proposed, e.g., He et al. (2020), but not about generic videos.

8.7 More effective scene change algorithms

Existing video super-resolution methods rarely involve the videos with scene change. In practice, a video sequence usually has many different scenes. When we consider the problem of video super-resolution on such videos, they have to be split into multiple segments without scene change and processed individually. This may result in large computational time. In fact, a simple subnet in 3DSRnet (Kim et al. 2019) has been proposed to deal with scene change, and it includes scene boundary detection and frame replacement. More dedicated networks that can process videos with complex scene changes are necessary for real-world applications.

8.8 More reasonable evaluation criteria for video quality

The criteria for evaluating the quality of super-resolution results mainly include PSNR and SSIM. However, their values are not able to reflect the video quality for human perception. That is, even if the PSNR value of a recovered video is high, the video also makes people uncomfortable. Therefore, new evaluation criteria for videos that are consistent with human perception need to be developed. More attentions have been attracted to the quality evaluation for images, such as Gu et al. (2020). However, the video quality including coherence between frames will be investigated in the future.

8.9 More effective methods for leveraging information

An important characteristic of video super-resolution methods is leveraging the information contained in video frames. The effectiveness of utilization influences the performance directly. Although many methods have been proposed, as mentioned in this paper, there are still some disadvantages. For instance, 3D convolution and non-local modules require a large amount of computation, and the accuracy of optical estimation can not be guaranteed. Therefore, the methods that can effectively utilize information contained in different frames is worth further studying.

9 Conclusions

In this survey, we reviewed the development of deep learning approaches for video super-resolution in recent years. We first classified existing video super-resolution algorithms into seven subcategories by the way of leveraging information contained in video frames, described the key ideas of representative methods and summarized the advantages and disadvantages of each method. Furthermore, we also compared and analyzed the performance of those methods on benchmark datasets, and outlines the wide applications of video super-resolution algorithms. Although the deep learning based VSR methods have made great progress, we listed eight open issues for the development of VSR algorithms, which is expected to provide some enlightment for researchers.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s10462-022-10147-y.

Acknowledgements We thank all the reviewers for their valuable comments. And we would like to thank Mr. Zekun Li (Master student at School of Artificial Intelligence in Xidian University) and Dr. Yaowei

Wang (Associate Professor with Peng Cheng Laboratory, Shenzhen, China) for their help in improving the quality of the manuscript. This work was supported by the National Natural Science Foundation of China (Nos. 61976164, 61876220, 61876221, and 61906184).

References

Bao W, Lai W, Zhang X, Gao Z, Yang M (2021) MEMC-Net: motion estimation and motion compensation driven neural network for video interpolation and enhancement. IEEE Trans Pattern Anal Mach Intell 43(3):933–948

Bare B, Yan B, Ma C, Li K (2019) Real-time video super-resolution via motion convolution kernel estimation. Neurocomputing 367:236–245

Brox T, Bruhn A, Papenberg N, Weickert J (2004) High accuracy optical flow estimation based on a theory for warping. In: Pajdla T, Matas J (eds) European conference on computer vision, pp 25–36

Burns C, Plyer A, Champagnat F (2017) Texture super-resolution for 3D reconstruction. In: Proceedings of the IAPR international conference on machine vision applications, pp 350–353

Caballero J, Ledig C, Aitken A, Acosta A, Totz J, Wang Z, Shi W (2017) Real-time video super-resolution with spatio-temporal networks and motion compensation. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 2848–2857

Chan KC, Wang X, Xu X, Gu J, Loy CC (2021a) GLEAN: generative latent bank for large-factor image super-resolution. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 14245–14254

Chan KC, Wang X, Yu K, Dong C, Loy CC (2021b) BasicVSR: the search for essential components in video super-resolution and beyond. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 4947–4956

Chan KC, Wang X, Yu K, Dong C, Loy CC (2021c) Understanding deformable alignment in video super-resolution. Proceedings of the AAAI conference on artificial intelligence, vol 35, pp 973–981

Chan KCK, Zhou S, Xu X, Loy CC (2021d) BasicVSR++: improving video super-resolution with enhanced propagation and alignment. arXiv preprint arXiv:2104.13371

Chen J, Tan X, Shan C, Liu S, Chen Z (2020) VESR-Net: the winning solution to Youku video enhancement and super-resolution challenge. arXiv preprint arXiv:2003.02115

Chen Y, Liu S, Wang X (2021) Learning continuous image representation with local implicit image function. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 8628–8638

Chu M, Xie Y, Mayer J, Leal-Taixé L, Thuerey N (2020) Learning temporal coherence via self-supervision for GAN-based video generation. ACM Trans Graph 39(4):75

Dai Q, Yoo S, Kappeler A, Katsaggelos AK (2015) Dictionary-based multiple frame video super-resolution. In: Proceedings of the IEEE international conference on image processing, pp 83–87

Dai J, Qi H, Xiong Y, Li Y, Zhang G, Hu H, Wei Y (2017) Deformable convolutional networks. In: Proceedings of the IEEE international conference on computer vision, pp 764–773

Daithankar MV, Ruikar SD (2020) Video super resolution: a review. In: ICDSMLA 2019, pp 488–495

Dario F, Huang Z, Gu S, Radu T et al (2020) Aim 2020 challenge on video extreme super-resolution: methods and results. arXiv preprint arXiv:2007.11803

Dasari M, Bhattacharya A, Vargas S, Sahu P, Balasubramanian A, Das SR (2020) Streaming 360-degree videos using super-resolution. In: Proceedings of the IEEE conference on computer communications, pp 1977–1986

Deshmukh AB, Rani NU (2019) Fractional-grey wolf optimizer-based kernel weighted regression model for multi-view face video super resolution. Int J Mach Learn Cybern 10(5):859–877

Dong C, Loy CC, He K, Tang X (2014) Learning a deep convolutional network for image super-resolution. In: European conference on computer vision, pp 184–199

Dong C, Loy CC, Tang X (2016) Accelerating the super-resolution convolutional neural network. In: European conference on computer vision, pp 391–407

Dosovitskiy A, Fischer P, Ilg E, Hausser P, Hazirbas C, Golkov V, Van Der Smagt P, Cremers D, Brox T (2015) FlowNet: learning optical flow with convolutional networks. In: Proceedings of the IEEE international conference on computer vision, pp 2758–2766

Drulea M, Nedevschi S (2011) Total variation regularization of local-global optical flow. In: 2011 14th IEEE international conference on intelligent transportation systems (ITSC), pp 318–323

Fakour-Sevom V, Guldogan E, Kämäräinen JK (2018) 360 panorama super-resolution using deep convolutional networks. In: International conference on computer vision theory and applications (VISAPP), pp 159–165

Farsiu S, Robinson MD, Elad M, Milanfar P (2004) Fast and robust multiframe super resolution. IEEE Trans Image Process 13(10):1327–1344

Fuoli D, Gu S, Timofte R (2019a) Efficient video super-resolution through recurrent latent space propagation. In: Proceedings of the IEEE/CVF international conference on computer vision workshop, pp 3476–3485

Fuoli D, Gu S, Timofte R, Tao X, Li W, Guo T, Deng Z, Lu L, Dai T, Shen X et al (2019b) Aim 2019 challenge on video extreme super-resolution: methods and results. In: Proceedings of the IEEE/CVF international conference on computer vision workshop, pp 3467–3475

Fuoli D, Huang Z, Gu S, Timofte R, Raventos A, Esfandiari A, Karout S, Xu X, Li X, Xiong X et al (2020) Aim 2020 challenge on video extreme super-resolution: methods and results. In: European conference on computer vision, pp 57–81

Gautam A, Singh S (2020) A comparative analysis of deep learning based super-resolution techniques for thermal videos. In: Proceedings of the international conference on smart systems and inventive technology, pp 919–925

Glaister J, Chan C, Frankovich M, Tang A, Wong A (2011) Hybrid video compression using selective keyframe identification and patch-based super-resolution. In: Proceedings of the IEEE international symposium on multimedia, pp 105–110

Gu J, Cai H, Chen H, Ye X, Ren J, Dong C (2020) Image quality assessment for perceptual image restoration: a new dataset, benchmark and metric. arXiv preprint arXiv:2011.15002

Guo J, Chao H (2017) Building an end-to-end spatial-temporal convolutional network for video super-resolution. In: Proceedings of the AAAI conference on artificial intelligence, pp 4053–4060

Guo K, Guo H, Ren S, Zhang J, Li X (2020) Towards efficient motion-blurred public security video super-resolution based on back-projection networks. J Netw Comput Appl 166:102691

Haris M, Shakhnarovich G, Ukita N (2018) Deep back-projection networks for super-resolution. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 1664–1673

Haris M, Shakhnarovich G, Ukita N (2019) Recurrent back-projection network for video super-resolution. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 3892–3901

Haris M, Shakhnarovich G, Ukita N (2020) Space-time-aware multi-resolution video enhancement. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 2859–2868

He K, Zhang X, Ren S, Sun J (2016) Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 770–778

He Z, He D, Li X, Xu J (2020) Unsupervised video satellite super-resolution by using only a single video. IEEE Geosci Remote Sens Lett 19:1–5

Hochreiter S, Schmidhuber J (1997) Long short-term memory. Neural Comput 9(8):1735–1780

Hu X, Mu H, Zhang X, Wang Z, Tan T, Sun J (2019) Meta-SR: a magnification-arbitrary network for super-resolution. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 1575–1584

Huang Y, Wang W, Wang L (2015) Bidirectional recurrent convolutional networks for multi-frame super-resolution. Adv Neural Inf Process Syst 28:235–243

Huang G, Liu Z, Van Der Maaten L, Weinberger KQ (2017) Densely connected convolutional networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 2261–2269

Huang Y, Wang W, Wang L (2018) Video super-resolution via bidirectional recurrent convolutional networks. IEEE Trans Pattern Anal Mach Intell 40(4):1015–1028

Hui T, Tang X, Loy CC (2018) LiteFlowNet: a lightweight convolutional neural network for optical flow estimation. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 8981–8989

Hui T, Tang X, Loy CC (2021a) A lightweight optical flow CNN-revisiting data fidelity and regularization. IEEE Trans Pattern Anal Mach Intell 43(8):2555–2569

Hui Z, Li J, Gao X, Wang X (2021b) Progressive perception-oriented network for single image super-resolution. Inf Sci 546:769–786

Ignatov A, Romero A, Kim H, Timofte R et al (2021) Real-time video super-resolution on smartphones with deep learning, mobile AI 2021 challenge: report. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops, pp 2535–2544

Ilg E, Mayer N, Saikia T, Keuper M, Dosovitskiy A, Brox T (2017) FlowNet 2.0: evolution of optical flow estimation with deep networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 1647–1655

Irani M, Peleg S (1991) Improving resolution by image registration. CVGIP Graph Models Image Process 53(3):231–239
Irani M, Peleg S (1993) Motion analysis for image enhancement: resolution, occlusion, and transparency. J Vis Commun Image Represent 4(4):324–335
Isobe T, Jia X, Gu S, Li S, Wang S, Tian Q (2020) Video super-resolution with recurrent structure-detail network. In: European conference on computer vision, pp 645–660
Jacobsen JH, Smeulders AW, Oyallon E (2018) i-RevNet: deep invertible networks. In: Proceedings of the international conference on learning representations
Jaderberg M, Simonyan K, Zisserman A, Kavukcuoglu K (2015) Spatial transformer networks. Adv Neural Inf Process Syst 28:2017–2025
Ji S, Xu W, Yang M, Yu K (2013) 3D convolutional neural networks for human action recognition. IEEE Trans Pattern Anal Mach Intell 35(1):221–231
Jia X, De Brabandere B, Tuytelaars T, Gool LV (2016) Dynamic filter networks. Adv Neural Inf Process Syst 29:667–675
Jiang K, Wang Z, Yi P, Jiang J (2018a) A progressively enhanced network for video satellite imagery superresolution. IEEE Signal Process Lett 25(11):1630–1634
Jiang K, Wang Z, Yi P, Jiang J, Xiao J, Yao Y (2018b) Deep distillation recursive network for remote sensing imagery super-resolution. Remote Sens 10(11):1700
Jo Y, Oh SW, Kang J, Kim SJ (2018) Deep video super-resolution network using dynamic upsampling filters without explicit motion compensation. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 3224–3232
Kalarot R, Porikli F (2019) MultiBoot VSR: multi-stage multi-reference bootstrapping for video super-resolution. In: Proceedings of the IEEE conference on computer vision and pattern recognition workshops, pp 2060–2069
Kappeler A, Yoo S, Dai Q, Katsaggelos AK (2016) Video super-resolution with convolutional neural networks. IEEE Trans Comput Imaging 2(2):109–122
Kim J, Lee JK, Lee KM (2016) Accurate image super-resolution using very deep convolutional networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 1646–1654
Kim TH, Sajjadi MSM, Hirsch M, Schölkopf B (2018a) Spatio-temporal transformer network for video restoration. In: European conference on computer vision, pp 111–127
Kim Y, Choi JS, Kim M (2018b) 2x super-resolution hardware using edge-orientation-based linear mapping for real-time 4K UHD 60 fps video applications. IEEE Trans Circuits Syst Express Briefs 65(9):1274–1278
Kim Y, Choi JS, Kim M (2018c) A real-time convolutional neural network for super-resolution on FPGA with applications to 4K UHD 60 fps video services. IEEE Trans Circuits Syst Video Technol 29(8):2521–2534
Kim SY, Lim J, Na T, Kim M (2019) Video super-resolution based on 3D-CNNS with consideration of scene change. In: Proceedings of the IEEE international conference on image processing, pp 2831–2835
Kupyn O, Budzan V, Mykhailych M, Mishkin D, Matas J (2018) Deblurgan: blind motion deblurring using conditional adversarial networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 8183–8192
Kwasniewska A, Ruminski J, Szankin M (2019) Improving accuracy of contactless respiratory rate estimation by enhancing thermal sequences with deep neural networks. Appl Sci 9(20):4405
Ledig C, Theis L, Huszr F, Caballero J, Cunningham A, Acosta A, Aitken A, Tejani A, Totz J, Wang Z, Shi W (2017) Photo-realistic single image super-resolution using a generative adversarial network. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 105–114
Lee Y, Yun J, Hong Y, Lee J, Jeon M (2018) Accurate license plate recognition and super-resolution using a generative adversarial networks on traffic surveillance video. In: Proceedings of the IEEE international conference on consumer electronics-Asia, ICCE-Asia, pp 1–4
Lei P, Todorovic S (2018) Temporal deformable residual networks for action segmentation in videos. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 6742–6751
Li Y, Li X, Fu Z, Zhong W (2016) Multiview video super-resolution via information extraction and merging. In: Proceedings of the 24th ACM international conference on multimedia, pp 446–450
Li K, Bare B, Yan B, Feng B, Yao C (2018) Face hallucination based on key parts enhancement. In: Proceedings of the IEEE international conference on acoustics, speech, and signal processing, pp 1378–1382
Li D, Liu Y, Wang Z (2019a) Video super-resolution using non-simultaneous fully recurrent convolutional network. IEEE Trans Image Process 28(3):1342–1355

Li S, He F, Du B, Zhang L, Xu Y, Tao D (2019b) Fast spatio-temporal residual network for video super-resolution. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 10522–10531
Li Y, Tsiminaki V, Timofte R, Pollefeys M, Gool LV (2019c) 3D appearance super-resolution with deep learning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 9671–9680
Li W, Tao X, Guo T, Qi L, Lu J, Jia J (2020) MuCAN: multi-correspondence aggregation network for video super-resolution. In: European conference on computer vision, pp 335–351
Liao R, Tao X, Li R, Ma Z, Jia J (2015) Video super-resolution via deep draft-ensemble learning. In: Proceedings of the IEEE international conference on computer vision, pp 531–539
Lim B, Son S, Kim H, Nah S, Lee KM (2017) Enhanced deep residual networks for single image super-resolution. In: Proceedings of the IEEE conference on computer vision and pattern recognition workshops, pp 1132–1140
Lin JY, Chang YC, Hsu WH (2020) Efficient and phase-aware video super-resolution for cardiac MRI. In: International conference on medical image computing and computer-assisted intervention (MICCAI), pp 66–76
Liu C, Sun D (2014) On Bayesian adaptive video super resolution. IEEE Trans Pattern Anal Mach Intell 36(2):346–360
Liu Z, Cui C (2018) A new low bit-rate coding scheme for ultra high definition video based on super-resolution reconstruction. In: Proceedings of the IEEE international conference on computer and communication technology, pp 325–329
Liu D, Wang Z, Fan Y, Liu X, Wang Z, Chang S, Huang T (2017) Robust video super-resolution with learned temporal dynamics. In: Proceedings of the IEEE international conference on computer vision, pp 2526–2534
Liu H, Gu Y, Wang T, Li S (2020a) Satellite video super-resolution based on adaptively spatiotemporal neighbors and nonlocal similarity regularization. IEEE Trans Geosci Remote Sens 58(12):8372–8383
Liu H, Ruan Z, Fang C, Zhao P, Shang F, Liu Y, Wang L (2020b) A single frame and multi-frame joint network for 360-degree panorama video super-resolution. arXiv preprint arXiv:2008.10320
Liu X, Shi K, Wang Z, Chen J (2021a) Exploit camera raw data for video super-resolution via hidden Markov model inference. IEEE Trans Image Process 30:2127–2140
Liu H, Zhao P, Ruan Z, Shang F, Liu Y (2021b) Large motion video super-resolution with dual subnet and multi-stage communicated upsampling. In: Proceedings of the AAAI conference on artificial intelligence, pp 2127–2135
Loshchilov I, Hutter F (2017) SGDR: stochastic gradient descent with warm restarts. In: Proceedings of the international conference on learning representations (ICLR)
Lucas BD, Kanade T (1981) An iterative image registration technique with an application to stereo vision. In: Proceedings of the international joint conference on artificial intelligence, pp 674–679
Lucas A, Lopez-Tapia S, Molina R, Katsaggelos AK (2019) Generative adversarial networks and perceptual losses for video super-resolution. IEEE Trans Image Process 28(7):3312–3327
Luo Y, Zhou L, Wang S, Wang Z (2017) Video satellite imagery super resolution via convolutional neural networks. IEEE Geosci Remote Sens Lett 14(12):2398–2402
Ma Z, Liao R, Tao X, Xu L, Jia J, Wu E (2015) Handling motion blur in multi-frame super-resolution. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 5224–5232
Nah S, Baik S, Hong S, Moon G, Son S, Timofte R, Lee KM (2019a) NTIRE 2019 challenge on video deblurring and super-resolution: dataset and study. In: Proceedings of the IEEE conference on computer vision and pattern recognition workshops, pp 1996–2005
Nah S, Timofte R, Gu S, Baik S, Hong S et al (2019b) NTIRE 2019 challenge on video super-resolution: methods and results. In: Proceedings of the IEEE conference on computer vision and pattern recognition workshops, pp 1985–1995
Odille F, Bustin A, Chen B, Vuissoz PA, Felblinger J (2015) Motion-corrected, super-resolution reconstruction for high-resolution 3D cardiac cine MRI. In: International conference on medical image computing and computer-assisted intervention (MICCAI), pp 435–442
Pan J, Cheng S, Zhang J, Tang J (2020) Deep blind video super-resolution. arXiv preprint arXiv:2003.04716
Patti AJ, Sezan MI, Tekalp AM (1997) Superresolution video reconstruction with arbitrary sampling lattices and nonzero aperture time. IEEE Trans Image Process 6(8):1064–1076
Peng C, Lin WA, Liao H, Chellappa R, Zhou SK (2020) SAINT: spatially aware interpolation network for medical slice synthesis. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 7750–7759

Poot DH, Van Meir V, Sijbers J (2010) General and efficient super-resolution method for multi-slice MRI. In: International conference on medical image computing and computer-assisted intervention (MICCAI), pp 615–622
Protter M, Elad M, Takeda H, Milanfar P (2009) Generalizing the nonlocal-means to super-resolution reconstruction. IEEE Trans Image Process 18(1):36–51
Ranjan A, Black MJ (2017) Optical flow estimation using a spatial pyramid network. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 2720–2729
Ren S, Guo H, Guo K (2019) Towards efficient medical video super-resolution based on deep back-projection networks. In: Proceedings of the IEEE international conference on iThings/GreenCom/CPSCom/SmartData, pp 682–686
Ronneberger O, Fischer P, Brox T (2015) U-net: convolutional networks for biomedical image segmentation. In: International conference on medical image computing and computer-assisted intervention (MICCAI), pp 234–241
Sajjadi MSM, Vemulapalli R, Brown M (2018) Frame-recurrent video super-resolution. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 6626–6634
Schultz RR, Stevenson RL (1996) Extraction of high-resolution frames from video sequences. IEEE Trans Image Process 5(6):996–1011
Shamsolmoali P, Zareapoor M, Jain DK, Jain VK, Yang J (2019) Deep convolution network for surveillance records super-resolution. Multimedia Tools Appl 78(17):23815–23829
Shi X, Chen Z, Wang H, Yeung DY, Wong Wk, Woo Wc (2015) Convolutional LSTM network: a machine learning approach for precipitation nowcasting. Adv Neural Inf Process Syst 28:802–810
Shi W, Caballero J, Huszr F, Totz J, Aitken AP, Bishop R, Rueckert D, Wang Z (2016) Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 1874–1883
Shocher A, Cohen N, Irani M (2018) Zero-shot super-resolution using deep internal learning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 3118–3126
Singh A, Singh J (2020) Survey on single image based super-resolution-implementation challenges and solutions. Multimed Tools Appl 79(3):1641–1672
Son S, Lee S, Nah S, Timofte R, Lee KM et al (2021) Ntire 2021 challenge on video super-resolution. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition Workshops, pp 166–181
Sun D, Yang X, Liu M, Kautz J (2018) PWC-Net: CNNs for optical flow using pyramid, warping, and cost volume. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 8934–8943
Sun W, Sun J, Zhu Y, Zhang Y (2020) Video super-resolution via dense non-local spatial-temporal convolutional network. Neurocomputing 403:1–12
Takeda H, Milanfar P, Protter M, Elad M (2009) Super-resolution without explicit subpixel motion estimation. IEEE Trans Image Process 18(9):1958–1975
Tao X, Gao H, Liao R, Wang J, Jia J (2017) Detail-revealing deep video super-resolution. In: Proceedings of the IEEE international conference on computer vision, pp 4482–4490
Tian Y, Zhang Y, Fu Y, Xu C (2020) TDAN: temporally-deformable alignment network for video super-resolution. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 3360–3369
Timofte R, De Smet V, Van Gool L (2014) A+: adjusted anchored neighborhood regression for fast super-resolution. In: Proceedings of the Asian conference on computer vision, pp 111–126
Tran D, Bourdev L, Fergus R, Torresani L, Paluri M (2015) Learning spatiotemporal features with 3D convolutional networks. In: Proceedings of the IEEE international conference on computer vision, pp 4489–4497
Umeda S, Yano N, Watanabe H, Ikai T, Chujoh T, Ito N (2018) HDR video super-resolution for future video coding. In: International workshop on advanced image technology, pp 1–4
Wang X, Girshick R, Gupta A, He K (2018) Non-local neural networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 7794–7803
Wang H, Su D, Liu C, Jin L, Sun X, Peng X (2019a) Deformable non-local network for video super-resolution. IEEE Access 7:177734–177744
Wang Z, Yi P, Jiang K, Jiang J, Han Z, Lu T, Ma J (2019b) Multi-memory convolutional neural network for video super-resolution. IEEE Trans Image Process 28(5):2530–2544
Wang X, Chan KCK, Yu K, Dong C, Loy CC (2019c) EDVR: video restoration with enhanced deformable convolutional networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition Workshops, pp 1954–1963

Wang L, Guo Y, Lin Z, Deng X, An W (2019d) Learning for video super-resolution through HR optical flow estimation. In: Proceedings of the Asian conference on computer vision, pp 514–529
Wang Z, Chen J, Hoi SC (2021a) Deep learning for image super-resolution: a survey. IEEE Trans Pattern Anal Mach Intell 43(10):3365–3387
Wang L, Wang Y, Lin Z, Yang J, An W, Guo Y (2021b) Learning a single network for scale-arbitrary super-resolution. In: Proceedings of the IEEE/CVF international conference on computer vision, pp 4801–4810
Wei Y, Chen L, Xie R, Song L, Zhang X, Gao Z (2019) FPGA based video transcoding system with 2K-4k super-resolution conversion. In: Proceedings of the IEEE international conference on visual communications and image processing, pp 1–2
Xiao A, Wang Z, Wang L, Ren Y (2018) Super-resolution for Jilin-1 satellite video imagery via a convolutional network. Sensors 18(4):1194
Xiao Z, Fu X, Huang J, Cheng Z, Xiong Z (2021) Space-time distillation for video super-resolution. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 2113–2122
Xin J, Wang N, Li J, Gao X, Li Z (2020) Video face super-resolution with motion-adaptive feedback cell. Proc AAAI Conf Artif Intell 34(7):12468–12475
Xu L, Jia J, Matsushita Y (2012) Motion detail preserving optical flow estimation. IEEE Trans Pattern Anal Mach Intell 34(9):1744–1757
Xue T, Chen B, Wu J, Wei D, Freeman WT (2019) Video enhancement with task-oriented flow. Int J Comput Vis 127(8):1106–1125
Yan B, Lin C, Tan W (2019) Frame and feature-context video super-resolution. In: Proceedings of the AAAI conference on artificial intelligence, pp 5597–5604
Yang W, Zhang X, Tian Y, Wang W, Xue JH, Liao Q (2019) Deep learning for single image super-resolution: a brief review. IEEE Trans Multimedia 21(12):3106–3121
Yi P, Wang Z, Jiang K, Jiang J, Ma J (2019) Progressive fusion video super-resolution network via exploiting non-local spatio-temporal correlations. In: Proceedings of the IEEE international conference on computer vision, pp 3106–3115
Ying X, Wang L, Wang Y, Sheng W, An W, Guo Y (2020) Deformable 3D convolution for video super-resolution. arXiv preprint arXiv:200402803
Yu H, Liu D, Shi H, Yu H, Wang Z, Wang X, Cross B, Bramler M, Huang TS (2017) Computed tomography super-resolution using convolutional neural networks. In: Proceedings of the IEEE international conference on image processing, pp 3944–3948
Zhang Y, Wu G, Yap PT, Feng Q, Lian J, Chen W, Shen D (2012) Reconstruction of super-resolution lung 4D-CT using patch-based sparse representation. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 925–931
Zhang T, Gao K, Ni G, Fan G, Lu Y (2018a) Spatio-temporal super-resolution for multi-videos based on belief propagation. Signal Process Image Commun 68:1–12
Zhang Y, Li K, Li K, Wang L, Zhong B, Fu Y (2018b) Image super-resolution using very deep residual channel attention networks. In: European conference on computer vision, pp 294–310
Zhang Y, Tian Y, Kong Y, Zhong B, Fu Y (2018c) Residual dense network for image super-resolution. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 2472–2481
Zhang W, Li H, Li Y, Liu H, Chen Y, Ding X (2021) Application of deep learning algorithms in geotechnical engineering: a short critical review. Artif Intell Rev 54:5633–5673
Zhu X, Hu H, Lin S, Dai J (2019a) Deformable ConvNets V2: more deformable, better results. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 9300–9308
Zhu X, Li Z, Zhang X, Li C, Liu Y, Xue Z (2019b) Residual invertible spatio-temporal network for video super-resolution. In: Proceedings of the AAAI conference on artificial intelligence, pp 5981–5988