

Optical Flow for Video Super-Resolution: A Survey

Zhigang Tu¹ · Hongyan Li^{2*} · Wei Xie^{3*} ·
Yuanzhong Liu¹ · Shifu Zhang⁴ · Baoxin Li⁵ ·
Junsong Yuan⁶

Received: date / Accepted: date

Abstract Video super-resolution is currently one of the most active research topics in computer vision as it plays an important role in many visual applications. Generally, video super-resolution contains a significant component, i.e., motion compensation, which is used to estimate the displacement between successive video frames for temporal alignment. Optical flow, which can supply dense and sub-pixel motion between consecutive frames, is among the most common ways for this task. To obtain a good understanding of the effect that optical flow acts in video super-resolution, in this work, we conduct a comprehensive review on this subject for the first time. This investigation covers the following major topics: the function of super-resolution (i.e., why we require super-resolution); the concept of video super-resolution (i.e., what is video super-resolution); the description of evaluation metrics (i.e., how (video) super-resolution performs); the introduction of optical flow based video super-resolution; the investigation of using optical flow to capture temporal dependency for video super-resolution. Prominently, we give an in-depth study of the deep learning based video super-resolution method, where some representative algorithms are analyzed

Zhigang Tu and Yuanzhong Liu
State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing, Wuhan University

E-mail: tuzhigang@whu.edu.cn

Corresponding author*: Hongyan Li
School of Information Engineering, Hubei University of Economics, Wuhan
E-mail: hongyanli2000@126.com.

Corresponding author*: Wei Xie
School of Computer, Central China Normal University, Wuhan
E-mail: XW@mail.ccnu.edu.cn.

Baoxin Li
School of Computing, Informatics, Decision System Engineering, Arizona State University
E-mail: baoxin.li@asu.edu

Junsong Yuan
Computer Science and Engineering department, State University of New York at Buffalo
E-mail: jsyuan@buffalo.edu

and compared. Additionally, we highlight some promising research directions and open issues that should be further addressed.

Keywords Video super-resolution · Optical flow · Optical Flow-based video super-resolution · Temporal dependency

1 Introduction

Image resolution reflects the visual details viewable in an image, thus typically, images with higher resolution capture more visual information than the low resolution ones for both human perception and machine interpretation (Park et al. 2003; Fookes et al. 2004; Nguyen et al. 2018). That is to say, higher resolution images supply clearer and more discriminative pictorial information for human to perceive, and finer details for machine to interpret (Xiong et al. 2010; Singh and Singh 2020). Consequently, capturing high resolution (HR) images is extremely important for both human and machines (Park et al. 2003; Ledig et al. 2017a).

However, low resolution (LR) images are widespread in the real world due to two main reasons related to the image sensors:

- The cost of an imaging sensor and its spatial resolution is directly related. Generally, an HR sensor is more costly than its LR counterpart, leading to the wide deployment of LR sensors in cost-sensitive applications.
- The spatial resolution is constrained by image sensing technology. Increasing the resolution may typically cause low signal-to-noise ratio (SNR) as less light can fall upon each sensor cell.

Given that many images are acquired by LR cameras, it is imperative to find a way for reconstructing HR images from captured LR images. This process is referred to as “super-resolution (SR) image reconstruction” (Farsiu et al. 2004; Shi et al. 2016). It has been an important and challenging technique in the field of computer vision and image processing, and continues to attract attention from the research community (Nasrollahi and Moeslund 2014; Thapa et al. 2016; Wang et al. 2021a).

Since the heuristic work of Tsai and Huang in 1984, SR has witnessed significant progresses (see Figure 1) while being widely used in many domains (Yuan et al. 2010; Borsoi et al. 2019):

- Visual entertainment: High-resolution images supply more comfortable visual experience, which is a long-term pursuit for human (Shen et al. 2015). Recently, with the emergence of new display technology, the high-definition television (HDTV, 1920×1080) and more advanced ultra high definition television UHD TV (3840×2048 or 4K, 7680×4320 or 8k), will dominate the consumer market. There is an increasing requirement for using SR method to transform LR videos into HR versions for being enjoyed on HR devices (Kappeler et al. 2016; Liu and Sun 2014; Liu et al. 2017).
- Video surveillance: Super-resolution is desirably required in video surveillance as it can supply more powerful and more distinguishable visual cues for many visual tasks, e.g., face recognition (Mudunuri and Biswas 2016; Chen et al. 2018;

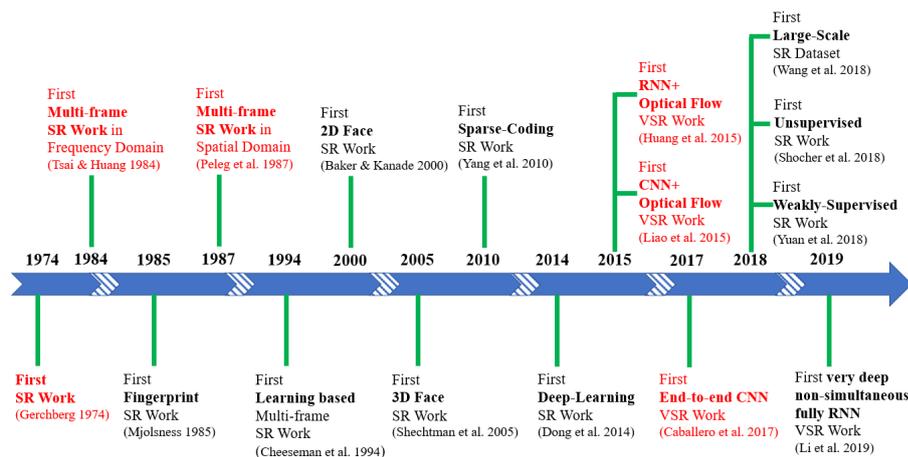


Fig. 1 Timeline of the landmark work in the field of super-resolution.

Ma et al. 2020), action recognition (Zhang et al. 2019), pose estimation (Hong et al. 2015; Sun et al. 2019), human activity recognition (Ryoo et al. 2017), person re-identification (Jing et al. 2017).

- Multimedia image processing: Super-resolution is effective for improving image and scene details, thus it is widely used in multimedia image processing, e.g., image denoising (Cruz et al. 2018; Huang et al. 2019), image inpainting (Meur et al. 2013), image retrieval (Tan et al. 2018), image classification (Cai et al. 2019), semantic segmentation (Zhao et al. 2018; Wang et al. 2020a), object detection (Girshick et al. 2016; Na and Fox 2017), object recognition (Ekeren et al. 2010; Sajjadi et al. 2017; Noor et al. 2019).
- Other applications: Super-resolution plays an important role in some other domains, e.g., medical imaging (Huang et al. 2017; You et al. 2020), remote sensing image processing (Tatem et al. 2001; Lei et al. 2020), infrared imaging (Choi et al. 2011; Han et al. 2018), biometrics (Huang et al. 2003; Yuan et al. 2009; Bian et al. 2017).

The main contributions of this work are four-fold:

1. Although some review papers about single image super-resolution have been published, but in the past decades, few important survey work on VSR has come off the press. We overview the definition, the application, and the landmark work of VSR, particularly with an emphasis on the optical flow based VSR;
2. We discuss the role and performance of optical flow in VSR systematically and comprehensively for the first time, and explain its principle;
3. We classify the traditional and recent optical flow based VSR techniques into three categories, and investigate the current deep learning (+ optical flow) based VSR algorithms. The advantages and limitations of each technique are summarized;

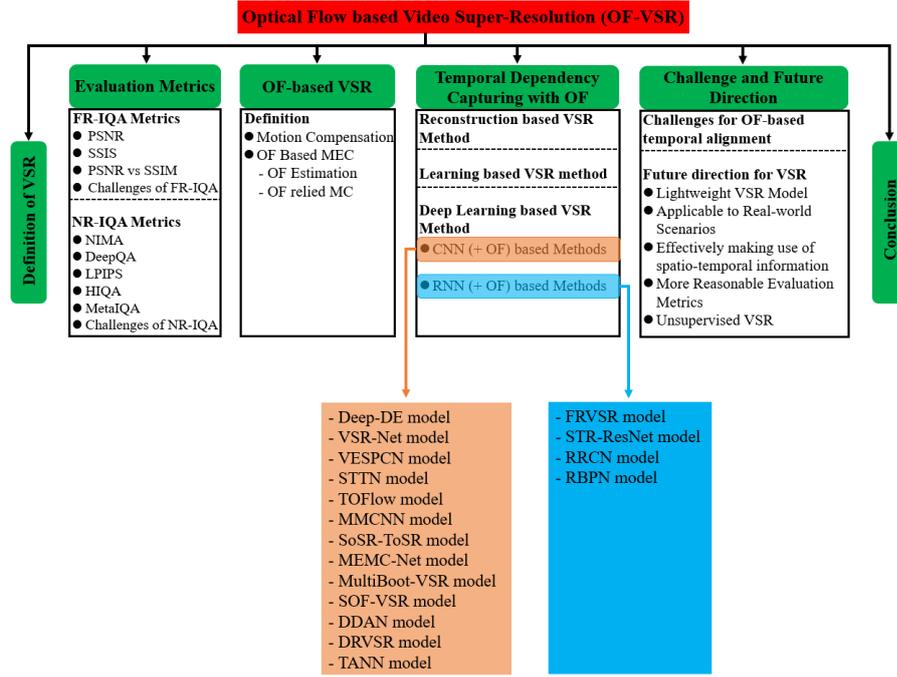


Fig. 2 A hierarchical-structured taxonomy about OF-VSR of this survey.

4. We analyze the challenges and open issues for both the optical flow based VSR and VSR. The new trends and promising directions are stated to supply an insightful guidance for the community.

The rest of the paper is organized as follows. We introduce the concept of video super-resolution in Section 2. Section 3 describes the optical flow based video super-resolution technique. The methods to capture the temporal dependency via optical flow for VSR are discussed in Section 4, with more attention given to deep learning (+ optical flow) methods. We present remaining challenges and future directions in Section 5. Finally, the conclusion is given in Section 6. Figure 2 shows the overall taxonomy about optical flow based video super-resolution (OF-VSR) covered in this review paper in a hierarchical-structured framework.

2 What is Video Super-Resolution

Video (or multi-frame) super-resolution (VSR) is a technique that addresses the issue of how to reconstruct high resolution (HR) images with better visual quality and finer spectral details by combining complementary information from multiple low-resolution (LR) counterparts (Peleg et al. 1987; Lin et al. 2005; Sajjadi et al. 2018; Daithankar and Ruikar 2020). VSR derives from a natural phenomenon that LR images are subsampled and contain sub-pixel shifts, and thus the complementary infor-

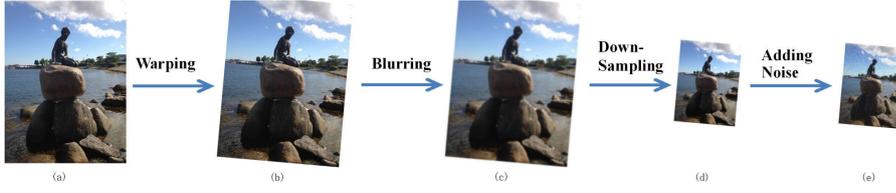


Fig. 3 A typical sample of the modeling of the degradation process. From left to right, the degradation process contains warping, blurring, downsampling, and noise. (a) is a ground-truth HR video frame, and (e) is the corresponding degraded LR video frame of (a). The sample is taken from [Nasrollahi and Moeslund 2014](#).

mation between LR images can be incorporated into a single image with higher resolution than the original observations ([Picku 2007](#); [Mudenagudi et al. 2011](#)). Enlarging the resolution can be treated as either improving the signal-to-noise ratio while preserving the size fixed, and/or depicting the image at a larger size with reasonable approximations for frequencies higher than those represented at the original size ([Lin et al. 2005](#); [Li et al. 2010](#)).

Mathematically, let \mathbf{I} represents a ground-truth HR video and \mathbf{I}_t represents a HR video frame at time t , \mathbf{F}_t denotes a degraded LR video frame. The degradation process of the HR video sequences can be expressed as follows:

$$\mathbf{F} = \mathbf{D}(\mathbf{I}; \eta) \quad (1)$$

where \mathbf{D} is the degradation function, η represents the parameters of the degradation function which includes various degradation factors, e.g. noise, motion blur, scaling factor.

In practice, the degradation process (i.e. \mathbf{D} and η) is unknown, as only the LR video frames \mathbf{F} are given, but the degradation factors, which are quite complicated, are unknown. Accordingly, we need to restore a HR video approximation $\hat{\mathbf{I}}$ of the ground truth HR video \mathbf{I} from the LR video \mathbf{F} , which can be formulated as:

$$\hat{\mathbf{I}} = \mathbf{S}(\mathbf{F}; \vartheta) \quad (2)$$

where \mathbf{S} denotes the VSR model and ϑ represents the parameters of the model \mathbf{S} .

The degradation process is usually affected by different factors (e.g., compression artifacts, anisotropic degradations, sensor noise, speckle noise) ([Wang et al. 2021a](#)). Many attempts have been tried to simulate this process ([Zhang et al. 2018a](#)), Figure 3 shows a typical example of the degradation process that is assumed in most SR methods. Currently, the widely adopted strategy is defined as:

$$\mathbf{D}(\mathbf{I}; \eta) = (\mathbf{F} \otimes k) \downarrow_s + n_\zeta, \{s, k, \zeta\} \subset \eta \quad (3)$$

where \downarrow_s represents a downsampling conduction which is performed with a scaling factor s , \otimes is a convolution operation and k is the blur kernel. n_ζ denotes some additive white Gaussian noise with a standard deviation ζ . ([Wang et al. 2021a](#)) stated that this combinative degradation model Eq.(3) is closer to realistic conditions and brings more benefits to VSR.

Finally, the objective function of VSR can be formulated as following:

$$\hat{\vartheta} = \arg_{\vartheta} \min L(\mathbf{I}, \hat{\mathbf{I}}) + \lambda \Phi(\vartheta) \quad (4)$$

where $L(\mathbf{I}, \hat{\mathbf{I}})$ denotes the loss function between the generated HR video $\hat{\mathbf{I}}$ and the ground truth video \mathbf{I} . $\Phi(\vartheta)$ represents the regularization term and λ is the tradeoff parameter.

From Eq. (4), we can find that VSR is an inversion problem as the process relies on the determination of the HR video \mathbf{I} from multiple low resolution observations \mathbf{F}_t .

Summary. Both spatial information with a frame and temporal information across different frames are important and useful for VSR (Li et al. 2016; Isobe et al. 2020).

3 Evaluation Metrics

Image quality assessment (IQA) refers to automatically evaluate the perceptual quality of a distorted image. IQA plays a crucial role in the low-level computer vision community due to it has a wide range of applications in image restoration, image retrieval, image quality monitoring systems, etc (Zhu et al. 2020; Zhai and Min 2020).

Since IQA normally serves as a basis for video quality assessment (VQA) because of videos are sequences of images, IQA is always the primary research topic. For understanding more evaluation metrics about IQA and VQA, please refer to a most recently survey paper (Zhai and Min 2020).

3.1 Full-Reference IQA (FR-IQA) Metrics

3.1.1 Peak-Signal-to-Noise Ratio (PSNR)

Peak-signal-to-noise ratio (PSNR), which is a crucial performance measure for lossy transformation, has long-term been widely used to evaluate the quality of both single image SR and VSR (Kim and Kwon 2010, Hore and Ziou 2010). For SR, PSNR is calculated in terms of the maximum pixel value and the mean squared error (MSE) between the ground truth image \mathbf{I}_{Gt} and the restored image \mathbf{I}_t :

$$PSNR = 10 \cdot \lg\left(\frac{M^2}{\frac{1}{N} \sum_{i=1}^N (\mathbf{I}_{Gt}(i) - \mathbf{I}_t(i))^2}\right) \quad (5)$$

where N represents the total number of pixels of the image *e.g.* \mathbf{I}_{Gt} , and M denotes the maximum pixel value (normally $M = 255$), i is a pixel position of the image. The MSE is computed as:

$$MSE = \frac{1}{N} \sum_{i=1}^N (\mathbf{I}_{Gt}(i) - \mathbf{I}_t(i))^2 \quad (6)$$

The PSNR value approaches infinity if $\mathbf{I}_{Gt} - \mathbf{I}_t = 0$, *i.e.* the super-resolved image \mathbf{I}_t is similar to the ground truth image \mathbf{I}_{Gt} . Which means that a higher PSNR value accompany with a higher reconstructed image quality.

Summary. The characteristics of PSNR can be summarized as following:

1. Depending on the pixel-level MSE and focuses on the corresponding pixels' difference;
2. Ignoring the human visual perception;
3. Insensitively to distinguish the structural content of image since different types of degradations can generate the same value of MSE;
4. Performing poor in complex real-world cases.

3.1.2 Structural Similarity Index Measure (SSIM)

To capture high quality perception, (Wang et al. 2004) proposed a structural similarity index measure (SSIM), which takes into account contrast, luminance distortion, and structure change between the two images. Since SSIM replaces the traditional error summation approaches (i.e. MSE) as PSNR with the structural similarity measuring, it can better simulate the human visual system (HVS) than PSNR. Specifically, SSIM is designed by integrating three factors :

$$SSIM = [L(\mathbf{I}_t, \mathbf{I}_{Gt})]^\alpha [C(\mathbf{I}_t, \mathbf{I}_{Gt})]^\beta [S(\mathbf{I}_t, \mathbf{I}_{Gt})]^\gamma \quad (7)$$

where the first factor $L(\mathbf{I}_t, \mathbf{I}_{Gt})$ is the luminance comparison function, the second factor $C(\mathbf{I}_t, \mathbf{I}_{Gt})$ is the contrast comparison function, and the third factor $S(\mathbf{I}_t, \mathbf{I}_{Gt})$ is the structure comparison function. α, β, γ are the weight parameters that used to adjust the relative importance of these three factors.

1) $L(\mathbf{I}_t, \mathbf{I}_{Gt})$ evaluates the closeness of the two images' mean luminance:

$$L(\mathbf{I}_t, \mathbf{I}_{Gt}) = \frac{2\mu_{\mathbf{I}_t}\mu_{\mathbf{I}_{Gt}} + C_1}{\mu_{\mathbf{I}_t}^2 + \mu_{\mathbf{I}_{Gt}}^2 + C_1} \quad (8)$$

where $\mu_{\mathbf{I}_t}$ and $\mu_{\mathbf{I}_{Gt}}$ respectively represents the mean luminance of \mathbf{I}_t and \mathbf{I}_{Gt} . For instance, $\mu_{\mathbf{I}_t}$ is computed as:

$$\mu_{\mathbf{I}_t} = \frac{1}{N} \sum_{i=1}^N \mathbf{I}_t(i) \quad (9)$$

$L(\mathbf{I}_t, \mathbf{I}_{Gt})$ is maximal and equal to 1 if $\mu_{\mathbf{I}_t} = \mu_{\mathbf{I}_{Gt}}$.

2) $C(\mathbf{I}_t, \mathbf{I}_{Gt})$ assesses the similarity of the two images' contrast.

$$C(\mathbf{I}_t, \mathbf{I}_{Gt}) = \frac{2\sigma_{\mathbf{I}_t}\sigma_{\mathbf{I}_{Gt}} + C_2}{\sigma_{\mathbf{I}_t}^2 + \sigma_{\mathbf{I}_{Gt}}^2 + C_2} \quad (10)$$

where $\sigma_{\mathbf{I}_t}$ and $\sigma_{\mathbf{I}_{Gt}}$ separately denotes the standard deviation of \mathbf{I}_t and \mathbf{I}_{Gt} . For example, $\sigma_{\mathbf{I}_t}$ is evaluated as:

$$\sigma_{\mathbf{I}_t} = \left[\frac{1}{N-1} \sum_{i=1}^N (\mathbf{I}_t(i) - \mu_{\mathbf{I}_t})^2 \right]^{\frac{1}{2}} \quad (11)$$

$C(\mathbf{I}_t, \mathbf{I}_{Gt})$ is maximal and equal to 1 when $\sigma_{\mathbf{I}_t} = \sigma_{\mathbf{I}_{Gt}}$.

3) $S(\mathbf{I}_t, \mathbf{I}_{Gt})$ is exploited to measure the correlation coefficient of the two images.

This is because the image structure can be expressed by the normalized pixel values (e.g. $\mathbf{I}_t - \mu_{\mathbf{I}_t}$), leading to their correlations are useful for reflecting the structural similarity. $S(\mathbf{I}_t, \mathbf{I}_{Gt})$ is calculated as:

$$S(\mathbf{I}_t, \mathbf{I}_{Gt}) = \frac{\sigma_{\mathbf{I}_t \mathbf{I}_{Gt}} + C_3}{\sigma_{\mathbf{I}_t} \sigma_{\mathbf{I}_{Gt}} + C_3} \quad (12)$$

where $\sigma_{\mathbf{I}_t \mathbf{I}_{Gt}}$ is expressed as:

$$\sigma_{\mathbf{I}_t \mathbf{I}_{Gt}} = \frac{1}{N-1} \sum_{i=1}^N (\mathbf{I}_t(i) - \mu_{\mathbf{I}_t})(\mathbf{I}_{Gt}(i) - \mu_{\mathbf{I}_{Gt}}) \quad (13)$$

C_1, C_2, C_3 are positive constants used for stability. $SSIM \in [0, 1]$, $SSIM = 0$ means there is no correlation between the two images, while $SSIM = 1$ means the super-resolved image \mathbf{I}_t equals to the ground-truth image \mathbf{I}_{Gt} .

Summary. The characteristics of SSIM can be summarized as following:

1. Suitable for evaluating the methods that without supplying sufficient texture details;
2. Preferring blur over texture mismatching.

3.1.3 PSNR vs SSIM

1. SSIM and PSNR are more sensitive to noise degradation than the other degradations;
2. PSNR is more sensitive to additive Gaussian noise, while SSIM is more sensitive to image compression.

3.1.4 Challenges of FR-IQA

1. FR-IQA measures unable to evaluate the perceptual quality precisely as they rely on the pixel-level error measures (e.g. L1 and L2 distances or their combination), leading to them focusing on pixel-level information locally;
2. FR-IQA measures lack of generalization as they are formulated with limited and refined constraints, and required artificial intervention, causing them usually fail to model unknown distortions;
3. FR-IQA measures are nearly unavailable in practice as they need non-distorted ground-truth reference images, but it is hard or impossible to obtain desired reference images in most cases.

3.2 No-Reference IQA (NR-IQA) Metrics: Deep Learning-Based Methods

3.2.1 NIMA

(Talebi and Milanfar 2018) explored a no-reference method, which called NIMA. It applies the CNN to predict the distribution of human opinion scores, leading to

it can be trained on both the aesthetic and pixel-level quality datasets. Importantly, the NIMA method predicts the distribution of ratings to a histogram, to replace the conventional approaches that classify images to low/high score or regress to the mean score. As a result, a high correlation to human perception is gained. The squared earth mover's distance (EMD) loss of (Hou et al. 2016) is selected, as it improves the classification performance with ordered classes.

The normalized EMD based loss function is expressed as:

$$EMD(p, \hat{p}) = \left(\frac{1}{N} \sum_{k=1}^N |\text{CDF}_p(k) - \text{CDF}_{\hat{p}}(k)|^r \right)^{1/r} \quad (14)$$

where $\text{CDF}_p(k)$ is a cumulative distribution function, which is computed as $\sum_{i=1}^k p_{S_i}$. Moreover, the predicted quality probabilities are input to a soft-max function to ensure $\sum_{i=1}^N \hat{p}_{S_i} = 1$. Same as (Hou et al. 2016), r is set to 2 for optimization with gradient descent more convenient.

3.2.2 DeepQA

Inspired by the prior CNN-based visual-task approaches to learn the deep feature map, (Kim and Lee 2017) proposed to use CNN to capture a visual sensitivity map, *i.e.* a weighting map of reflecting the visual importance of each pixel to HVS. Accordingly a DeepQA model is formed, which in fact a CNN based full-reference image quality assessment (FR-IQA) model. In contrast to the traditional IQA measures, DeepQA aims to learning the optimal visual weight from the IQA dataset itself without requiring any prior knowledge of the HVS, where the training process requires some information of the dataset: a triplet of distorted images, objective error maps, and the subjective scores. Specifically:

1) The objective error map is defined as:

$$e = \frac{\log(1/((\hat{\mathbf{I}}_R - \hat{\mathbf{I}}_D)^2 + (\varepsilon/255^2)))}{\log(255^2/\varepsilon)} \quad (15)$$

where \mathbf{I}_R and \mathbf{I}_D respectively denotes the reference image and the distorted image, $\hat{\mathbf{I}}_R$ and $\hat{\mathbf{I}}_D$ are their normalized version.

2) The visual sensitivity map learned from CNN is defined as:

$$s_1 = CCN_1(\hat{\mathbf{I}}_D; \theta_1) \quad (16)$$

$$s_2 = CCN_2(\hat{\mathbf{I}}_D, e; \theta_2) \quad (17)$$

where θ_1 and θ_2 are the parameters of DeepQA. The perceptual error map is estimated by:

$$p = s \odot e \quad (18)$$

where \odot is the Hadamard product, s is s_1 or s_2 .

Accordingly, the pooled score is computed by averaging the cropped perceptual error map:

$$\mu_p = \frac{1}{(H-8)(W-8)} \sum_{(i,j) \in \omega} p \quad (19)$$

where H and W are the height and width of p , (i, j) is pixel index, and ω denotes the cropped region.

3) The final objective function of the DeepQA model is expressed as:

$$\mathcal{L}_s(\hat{\mathbf{I}}_D; \theta) = \|(f(\mu_p) - S)\|_F^2 \quad (20)$$

where $f(\cdot)$ denotes a nonlinear regression function, and S represents the subjective score of the input distorted image.

The structure of DeepQA is described in Figure 4.

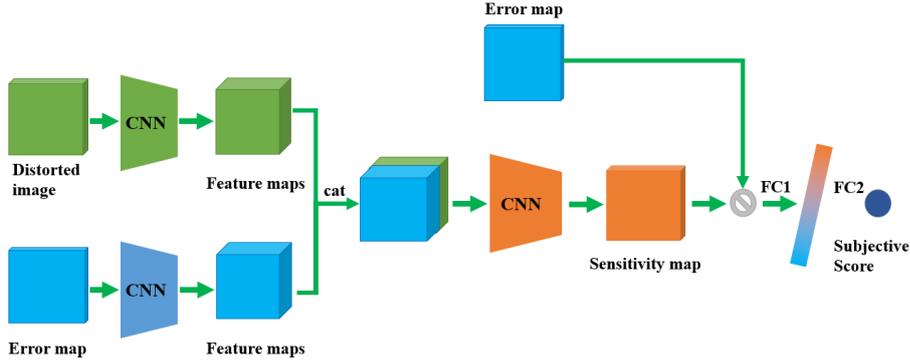


Fig. 4 The architecture of the DeepQA model. Input: A distorted image and An error map, then producing a sensitivity map. Output: a subjective score, which is regressed by multiplying with the error map.

Summary. The characteristics of DeepQA can be summarized as following:

- The DeepQA model, which with the usage of a triplet of a distorted image, its objective error map, and the subjective score, can capture the human visual sensitivity without any prior knowledge.
- A total variation regularization, which penalizes the high frequency components of the estimated sensitivity map, enables the sensitivity map to more visually plausible without reducing the performance.
- The model, which is optimized in an end-to-end manner, obtains the state-of-the-art correlation with human subjective scores.

3.2.3 LPIPS

(Zhang et al. 2018b) construct a large-scale dataset for human perceptual similarity evaluation. Importantly, they assess the perceptual image patch similarity (LPIPS) by comparing the deep features with the classical metrics, where the deep features are learnt by CNNs with different architectures and visual tasks. The experimental results reveal that the deep features can model the perceptual similarity much better than the prior metrics that without CNNs.

The key component of LPIPS is the *network activations to distance*, which can be expressed as:

$$D(p, p_0) = \sum_l \frac{1}{H_l W_l} \sum_{h,w} \|w_l \odot (\hat{F}_{hw}^l - \hat{F}_{0hw}^l)\|_2^2 \quad (21)$$

where $D(p, p_0)$ represents the cosine distance between the reference and the distorted patches p and p_0 . l denotes a convolution layer. w is a vector used to scale a channel. h represents the predict perceptual judgment. \hat{F} denotes the extracted deep feature, and $\hat{F}^l, \hat{F}_0^l \in \mathfrak{R}^{H_l \times W_l \times C_l}$ for layer l . (Zhang et al. 2018b) scale the activations channel-wise via vector $w^l \in \mathfrak{R}^{C_l}$ and calculate the l_2 distance.

Summary. The characteristics of LPIPS can be summarized as following:

- The stronger a feature set is at classification and detection, the stronger it is as a model of perceptual similarity judgments.
- A good feature is a good feature. Features that are good at semantic tasks also provide good models of both human perceptual behavior and macaque neural activity.

3.2.4 Hallucinated-IQA (HIQA)

(Lin and Wang 2018) presented a hallucination-guided quality regression network (named HIQA), which takes the perceptual discrepancy into a deep neural network for learning, to imitate HVS and defeat the ill-posed nature of NR-IQA. Specifically, by making use of the perceptual discrepancy between the distorted image and the hallucinated reference, HIQA achieves an accurate and robust perceptual prediction.

As shown in Figure 5, HIQA composes of three heavily related subnets.

A. Quality Aware Generative Network

The exploited Quality-Aware Generative Network aims to produce hallucinated reference images, where the hallucinated reference image is used to compensate the absence of true reference image. The gap between the hallucinated reference and the true reference is less, the performance of the quality regression network is better.

Importantly, to obtain high quality hallucinated reference images, HIQA exploits a quality-aware perceptual loss, which is able to incorporate the deep features of the regression network dynamically. The loss function is defined as:

$$L_s(G_\theta(I_d^i), I_r^i) = \lambda_1 L_v(G_\theta(I_d^i), I_r^i) + \lambda_2 L_q(G_\theta(I_d^i), I_r^i) \quad (22)$$

where

$$L_v = \sum_{C_v=1}^{C_v} \frac{1}{W_j H_j} \sum_{x=1}^{W_j} \sum_{y=1}^{H_j} \|\phi_j(G_\theta(I_d^i))_{x,y} - \phi_j(I_r^i)_{x,y}\|^2 \quad (23)$$

and

$$L_q = \sum_{C_q=1}^{C_q} \frac{1}{W_k H_k} \sum_{x=1}^{W_k} \sum_{y=1}^{H_k} \|\pi_k(G_\theta(I_d^i))_{x,y} - \pi_k(I_r^i)_{x,y}\|^2 \quad (24)$$

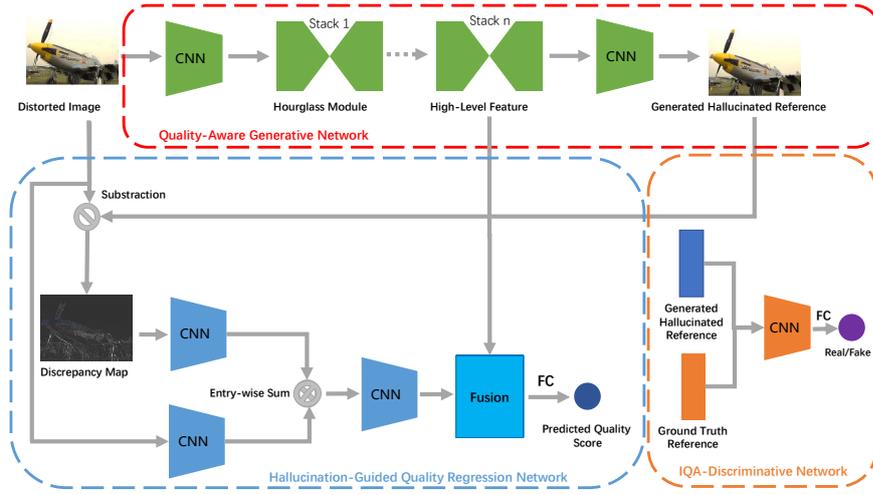


Fig. 5 The architecture of the Hallucinated-IQA (HIQA) framework. There are three main components: (a) A Quality-Aware Generative Network G , which is utilized to produce hallucinated reference images; (b) A Hallucination-Guided Quality Regression Network R , which is used to incorporate the discrepancy information between the hallucinated image and the distorted image encoded in the discrepancy map; (c) A IQA-Discriminator D , which is applied to refine the hallucinated image further.

Particularly, I_d is the distorted image, and $G(I_d)$ is the generating function; $I_d^i, \{i = 1, 2, \dots, N\}$ denotes a set of distorted images, and $I_r^i, \{i = 1, 2, \dots, N\}$ denotes the corresponding true reference images; $\phi_j(\cdot)$ represents the feature map at the j -th layer of VGG-19, and $\pi_k(\cdot)$ represents the feature map at the k -th layer of the hallucination-guided quality regression network R ; W and H denote the dimensions of the feature map, C is the number of feature maps at a particular layer.

B. IQA-Discriminative Network

IQA-Discriminator D aims to reduce the affect of bad hallucination images to the deep regression network R , which is realized by distinguishing the fake samples from the real samples based on their positive or negative impact to R .

G is optimized to fool the IQA-discriminator D by producing the qualified hallucination scene to help improve R , and the adversarial loss of G can be expressed as:

$$L_{adv} = \mathbf{E}[\log(1 - D_\omega(G_\theta(I_d)))] \quad (25)$$

The overall loss function of G for all training samples is defined by

$$L_G = \mu_1 L_p + \mu_2 L_s + \mu_3 L_{adv} \quad (26)$$

where μ_1, μ_2, μ_3 are the parameters which are aiming at keeping the trade off between the three loss components.

C. Hallucination Guided Quality Regression Network

The Hallucination-Guided Quality Regression Network R , which incorporates the discrepancy information with the high-level semantic fusion from the generative network G , to provide itself more plentiful and valid information. This is very useful to guide the network training.

Discrepancy Map. (Lin and Wang 2018) treat the distorted images and their discrepancy maps as pairs $\{I_d^i, I_{map}^i\}_{i=1}^N$, a deep regression network can be trained according to:

$$\hat{\gamma} = \arg \min_{\gamma} \frac{1}{N} \sum_{i=1}^N L_r(R(I_d^i, I_{map}^i), s^i) \quad (27)$$

where the discrepancy map is defined as:

$$I_{map} = |I_d - G_{\hat{\theta}}(I_d)| \quad (28)$$

High-level Semantic Fusion. The fusion term is defined as:

$$F = f(H_{5,2}(I_d)) \otimes (R_1(I_d, I_{map})) \quad (29)$$

where f is a linear projection to make the dimensions of H and R_1 equally, R_1 represents the feature extraction before the fully connected layers (R_2) of R , and \otimes represents concatenation.

The Overall Loss Function. The final loss of R is expressed as:

$$L_R = \frac{1}{T} \sum_{t=1}^T \|R_2(f(H_{5,2}(I_d))) \otimes R_1(I_d, I_{map}) - s^t\|_{\ell_1} \quad (30)$$

3.2.5 MetaIQA

To model the human beings' ability that "getting the quality prior knowledge from images with various distortions easily and adapting to evaluate unknown distorted images quickly", (Zhu et al. 2020) proposed a NR image quality metric MetaIQA based on the deep meta-learning, in which MetaIQA can capture the meta-knowledge shared by human when to assess the quality of images with multifarious distortions. In brief, the exploited MetaIQA enables the machines learn to learn, that is, to obtain the capacity of learning quickly from a relatively small amount of training samples for a related new task. Figure 6 shows the entire procedure of MetaIQA which is summarized in Algorithm 1.

3.2.6 Challenges of NR-IQA

1. **The definition is ill-posed:** NR-IQA generally takes the distorted image as input for evaluation without any additional data. However, this scheme is counter-intuitive, as HVS requires a reference to measure the perceptual discrepancy by comparing the distorted image either directly with the original undistorted image or implicitly with a hallucinated scene in mind (Lin and Wang 2018);
2. **The generalization ability is limited:** The deep-learning based NR-IQA metrics usually depend on pre-trained networks, however, these pre-trained networks are not designed for IQA, leading to their generalization performance is unsatisfactory when assessing unknown types of distortions.

Algorithm 1 Meta-learning based IQA (MetaIQA)

Input: Meta-training set $\mathcal{D}_{meta}^{p(\tau)} = \{\mathcal{D}_s^{\tau_i}, \mathcal{D}_q^{\tau_i}\}_{i=1}^N$, where $\mathcal{D}_{tr_q}^{\tau_i}$ and $\mathcal{D}_{tr_s}^{\tau_i}$ are task-support set and task-query set, and N is the total number of tasks, a target NR-IQA task with M training images, query image x , learning rate β

Output: Predicted quality score \hat{y} for x

- 1: Initialize model parameters θ ;
- 2: /* meta-training for prior model */
- 3: **for** $iteration = 1, 2, \dots$ **do**
- 4: Sample a mini-batch of k tasks in $\mathcal{D}_{meta}^{p(\tau)}$;
- 5: **for** $i = 1, 2, \dots, k$ **do**
- 6: /* first level computing */
- 7: Compute $\theta'_i = Adam(\mathcal{L}_{\tau_i}, \theta)$ on $\mathcal{D}_s^{\tau_i}$;
- 8: /* second level computing */
- 9: Compute $\theta_i = Adam(\mathcal{L}_{\tau_i}, \theta'_i)$ on $\mathcal{D}_q^{\tau_i}$;
- 10: **end for**
- 11: update $\theta \leftarrow \theta - \beta \frac{1}{k} \sum_{i=1}^k (\theta - \theta_i)$;
- 12: **end for**
- 13: /* fine-tuning for NR-IQA task */
- 14: Update $\theta_{te} = Adam(\mathcal{L}, \theta)$ on the NR-IQA task;
- 15: Input x into the quality model $f_{\theta_{te}}$;
- 16: **return** \hat{y} .

Fig. 6 The entire procedure of Meta-learning based IQA (MetaIQA). The procedure of Algorithm 1 is copied from (Zhu et al. 2020).

4 Optical Flow Based VSR

Generally, most of the VSR methods contain three basic components (Lin et al. 2005): (a) Motion compensation (alignment); (b) Interpolation; (c) Blur and noise removal (restoration).

4.1 Motion Compensation

The first component, i.e., motion compensation, which aims to obtain image alignment between multiple video frames, is a key component for VSR. Human vision system is sensitive to motion, thus capturing and modeling the influence of motion on visual perception is crucial for VSR (Li et al. 2016; Liu et al. 2018). To get accurate video super-resolution result, the motion between video frames should be estimated as accurate as possible, because inaccurate motion distorts local structure and degrades the final HR image reconstruction (Su et al. 2012; Liao et al. 2015). Consequently, how to perform a good motion estimation has attracted great attention.

Numerous motion estimation techniques, e.g., the global parametric models, and the local non-parametric models (e.g., block-motion approaches and optical flow based methods), have been proposed to improve the performance of video super-resolution (Schoenemann and Cremers 2012).

Furthermore, sub-pixel shifts is another key factor for VSR (Babacan et al. 2011). If the LR images only include integer shifts, there is no new information can be

produced when combining them to reconstruct the HR image. In other words, capturing sub-pixel shifts is necessary in motion estimation for VSR (Dai et al. 2017).

Since optical flow is able to supply accurate and sub-pixel motion information (Tu et al. 2014; Tu et al. 2019), the optical flow based VSR method has been studied for a long time, leading to significant progress in the past decade (Nguyen et al. 2018; Anwar et al. 2020). Optical flow can model the temporal dependency between consecutive video frames (Wang et al. 2020b), where the temporal dependency is normally considered as an essential component of VSR, and thus the estimation of optical flow can significantly affect the final result of VSR (Huang et al. 2018). Besides, the optical flow field is a dense motion field that can describe the deformation or mapping of every pixel between two video frames, therefore the optical flow based VSR is very suitable for extracting the mapping of non-rigid moving objects in the video, and thus contributing to addressing the challenge of super-resolution for non-rigid objects in VSR.

4.2 Optical Flow Based Motion Estimation and Compensation (MEC)

4.2.1 Optical Flow (Motion) Estimation

Optical flow estimation is based on the assumption that the brightness of a moving pixel remains constant over time. Mathematically, optical flow estimation is formulated as follows:

$$E(u, v) = F(I_t, I_{t+1}; \Theta_F) \quad (31)$$

where I_t and I_{t+1} are two successive input video frames, which separately denote the current frame at time t (i.e. the target frame) and the next frame at time $t + 1$ (i.e. the neighboring frame). E refers to the estimation operation of optical flow, F is a function utilized to calculate optical flow and Θ_F represents its parameters. $W = (u, v)$ denotes the calculated optical flow, with the horizontal and vertical flow components u and v respectively.

Figure 7 shows the visualized optical flow. As (Tu et al. 2019) stated: “The 2D displacement field, which describes the apparent motion of brightness patterns between two successive images, is called the optical flow.” “The optical flow field is ideally a dense field of displacement vectors (see Figure 7 (b), (c)), which maps all points of the first image onto their corresponding locations in the second image.” Particularly, Figure 7 (b) is the color-coded ground truth flow, and (Tu et al. 2019) explained: “The color-coded flow field is a dense visualization of the optical flow field. A color hue is associated to each direction and the saturation of the color increases with the magnitude of the flow vector.” Figure 7 (c) is the vector plot ground truth flow, and (Tu et al. 2019) described: “which directly represents the displacement vectors and provides a good intuitive perception of physical motion.” To understand more knowledge about optical flow, please refer to the optical flow survey paper of (Tu et al. 2019).

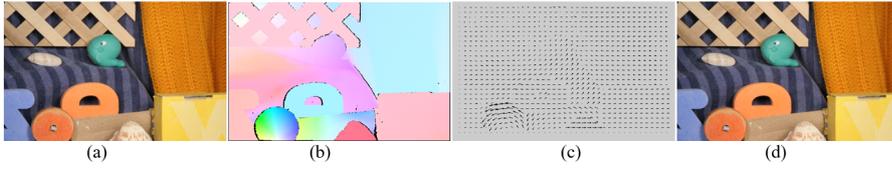


Fig. 7 Optical flow: (a) and (d) respectively represents frame 10 and frame 11 of the RubberWhale sequence on the Middlebury benchmark (Baker et al. 2011); (b) is the color-coded ground truth flow; (c) is the vector plot ground truth flow. The sample is taken from (Tu et al. 2019).

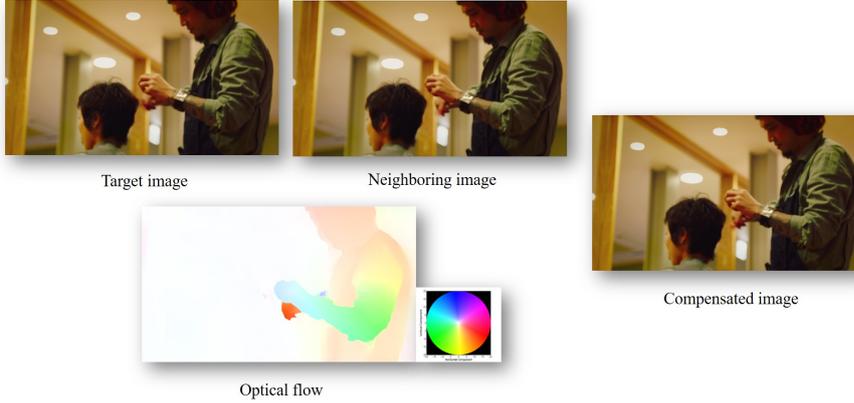


Fig. 8 The general flowchart of motion estimation and compensation (MEC) in VSR.

4.2.2 Optical Flow Relied Motion Compensation

As shown in Figure 8, motion compensation is generally conducted by performing image transformation between an image sequence according to the estimated motion information to make neighboring frames matching with the target frame spatially. In particular, this operation can be achieved by certain methods, e.g., the geometric registration (Nasrollahi and Moeslund 2014; Liu et al. 2020). Mathematically, a compensated frame is represented as:

$$I_{MC} = F_{MC}(I, W; \theta_{ME}) \quad (32)$$

where $F_{MC}(\cdot)$ denotes a motion compensation function, I is the neighboring frame, W represents the estimated optical flow, and θ_{ME} denotes the parameters of optical flow based motion estimation. Please refer to Figure 8 for motion estimation and motion compensation in detail.

Summary. Motion cue plays a crucial role in capturing temporal dependency between LR video frames for VSR:

- Motion compensation, which is one of the main components in VSR, encodes temporal dependency in compensated LR video frames in terms of estimating temporal information from consecutive frames;

- Optical flow, which supplies accurate, sub-pixel and dense motion information, as well as also can capture the motion of objects that are non-rigid, non-planar and self-occluded (Nasrollahi and Moeslund 2014), is good for modeling the temporal information.

5 Temporal Dependency Capturing for VSR with OF

In contrast to a single image, adjacent video frames provide temporal correlations. Therefore, to conduct VSR, it is essential to exploit temporal dependency between consecutive frames efficiently and effectively (Wang et al. 2020b). To model the temporal dependency, optical flow is extensively utilized (Caballero et al. 2017; Liu et al. 2018; Wang et al. 2020b). The VSR methods, which used optical flow to capture the temporal dependency, can be broadly classified into three main categories: (1) reconstruction based VSR method, (2) learning-based VSR method and, (3) deep learning based VSR method. While the first two approaches can be treated as the traditional VSR method, the deep learning based VSR method can be further classified into two groups: (a) the CNN (+optical flow) based VSR method and; (b) the RNN (+optical flow) based VSR method. Figure 9 shows the categories of optical flow based VSR methods. The first three VSR methods always have explicit motion compensation, while the RNN based VSR methods normally do not have explicit motion compensation. The first two traditional VSR methods have been studied for decades, but are surpassed by their deep learning based counterparts.

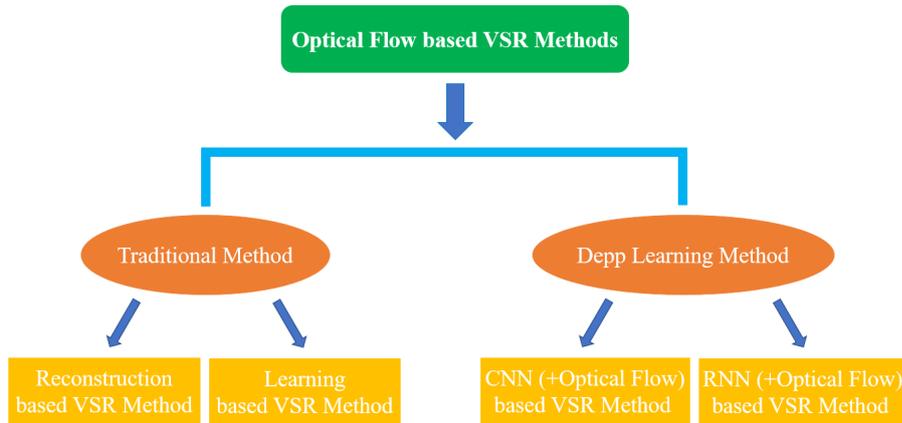


Fig. 9 The categories of optical flow based VSR methods.

Summary. Most of the optical flow related VSR methods exploit motion information in two main aspects:

- Explicitly registering multiple video frames in terms of the estimated motion;
- Implicitly embedding motion estimation to regularize the process of HR image reconstruction.

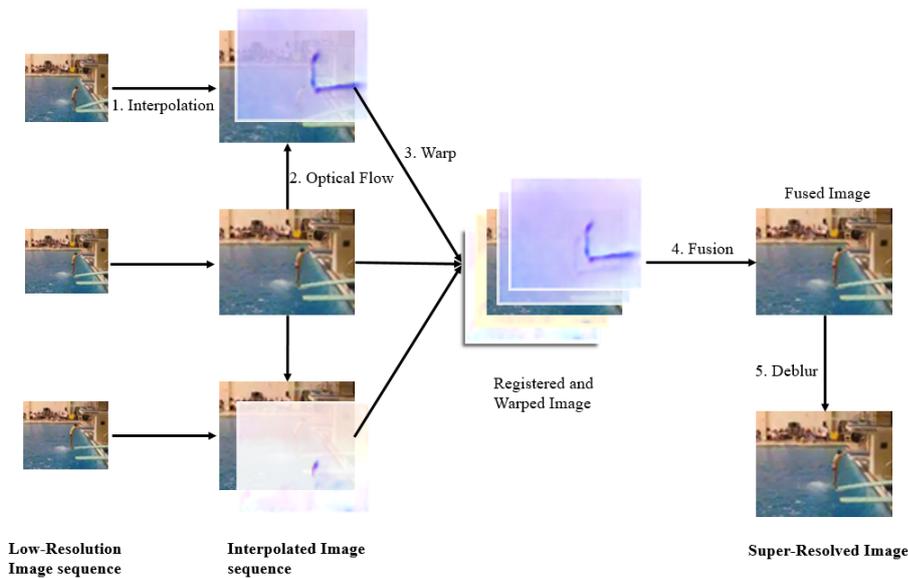


Fig. 10 The flow-chart of the Reconstruction (+ optical flow) based VSR method. This method usually includes 5 steps: 1) Performing interpolation to the LR image sequence to get higher resolution image sequence; 2) Computing optical flow between successive images; 3) Registering images to the reference image by using the motion information of the estimated optical flow; 4) Estimating the SR image by fusing the reference image and the registered images; 5) Recovering the final SR image via deblurring.

5.1 Reconstruction based VSR Method

To model the temporal dependency in VSR, some algorithms first perform optical flow estimation explicitly to compute sub-pixel motion between consecutive LR video frames and then warp each LR image to the target HR space according to the computed optical flow. In this way, the difference caused by the movement between the LR video frames can be captured, and the correspondence between the LR observations and the desired reconstructed HR image can be exploited, where they are useful for guiding the reconstruction of the targeted HR image. This kind of algorithms is reconstruction (+ optical flow) based VSR methods, which will be called reconstruction-based VSR methods for short (Mitzel et al. 2009; Huang et al. 2015; Liao et al. 2015).

For the reconstruction-based VSR methods involving optical flow estimation, since LR video frames have different sub-pixel motions and rotations from each other, it is crucial to obtain motion information precisely before fusing them to produce an HR image. Inaccurate motion cue will lead to various types of visual artifacts that subsequently damage the quality of the reconstructed HR image (Thapa et al. 2016). In brief, the accuracy and efficiency of optical flow estimation critically affects the performance of the reconstruction-based VSR. Figure 10 shows the framework of the Reconstruction (+ optical flow) based VSR method.

Baker and Kanade 1999 proposed a pioneer work to estimate optical flow for VSR to address the issue of complex motions in realistic videos. The correlation

between super-resolution optical flow and pyramid-based image representations is analyzed. [Zhao and Sawhney 2002](#) gave a systematic study to discover the impact of image alignment and warping errors on the VSR. Specifically, the performance of VSR under alignment with piecewise parametric or optical flow based approaches are investigated. They revealed that optical flow is especially significant for reconstructing high-frequency components in the HR image, and the flow consistency and flow accuracy are two critical elements. Optical flow consistency across video frames are important, and optical flow errors could cause super-resolution to become infeasible. [Mitzel et al. 2009](#) proposed a variational framework for VSR with arbitrary videos. This work contains two important steps: firstly, a quadratic relaxation strategy, which is able to compute high accuracy optical flow, is introduced to get motion for mapping LR images. Secondly, a variational method, which can impose a total variation regularity to the computed intensity map, is used to estimate the HR image. [Keller et al. 2011](#) exploited a full motion-compensated variational framework to design a simultaneous VSR method, which is able to jointly compute the HR image sequence and its corresponding HR optical flow. In this way, the VSR performance is boosted as more accurate details can be propagated from frame to frame. There are two main contributions of [Keller et al. 2011](#): (1) it is the first work to calculate super-resolved optical flows; (2) this method is possible to be used to general video with arbitrary scene content and/or arbitrary optical flows. In particular, the optical flow is estimated according to a classical total variational energy function ([Papenberg et al. 2006](#)).

On the other side, to alleviate the degradation caused by the estimated inaccurate dense optical flow in the reconstruction based VSR method, [Su et al. 2012](#) proposed to compute local flow with reliable accuracy based on the sparse feature point (e.g. SIFT, corner point) correspondences. This strategy is effective even when the input video frames are with noise, large-scale or other complex local motion.

Limitation. Reconstruction-based VSR methods with explicit optical flow estimation generally suffer from the following limitations:

- High computational cost for optical flow estimation;
- Difficulty in obtaining high quality motion information even with the state-of-the-art optical flow approaches;
- Visual artifacts inevitably caused by inaccurate registration of erroneous motions during the reconstruction;
- Degenerated cases: Inaccurate optical flow may sometimes lead a poor reconstruction performance worse than direct image interpolation.

5.2 Learning based VSR method

Reconstructing the HR image from LR video frames is an ill-posed problem and needs to be regularized ([Kappeler et al. 2016](#)). Consequently, some probabilistic models, e.g., Expectation-Maximization (EM) framework and Bayesian framework, are proposed to introduce priors to control the smoothness or the total variation of the image ([Mudenagudi et al. 2011](#)). This kind of VSR approaches are considered as the learning (+ optical flow) based VSR method, called learning based VSR method for short ([Liu and Sun 2014](#); [Ma et al. 2015](#)).

Generally, VSR methods make some simplifying assumptions. For example, the underlying motion may be assumed to have an oversimplified parametric form, or the blur kernel and noise levels are assumed to be known. But in practice, the movement of objects and cameras can be arbitrary, the motion blur and point spread functions can result in an unknown blur kernel, and the noise levels in video are unknown. Therefore, the learning based VSR methods attempt to integrate all these factors in a single framework without making oversimplified assumptions, and optimized them simultaneously.

Liu and Sun 2014 exploited a Bayesian method for adaptive VSR by estimating optical flow, blur kernel and noise level in addition to reconstruct the original HR video frames simultaneously in a single framework. They optimize the optical flow and the noise level jointly in a coarse-to-fine manner on a Gaussian image pyramid. At each pyramid level, the optical flow and noise level are computed iteratively in the maximum a posterior (MAP) way. To address the issue of motion blur in VSR, Ma et al. 2015 proposed to search least blurred pixels in VSR optimally. An EM framework was designed to guide residual blur estimation and HR image reconstruction. The classical optical flow regularizer (Sun et al. 2010) was selected for supplying the motion prior. To reduce the computational cost of motion estimation, they used the interpolated TV-L1 optical flow (Sun et al. 2010) on the LR images for approximation.

Limitation. The learning-based VSR methods usually formulate VSR as an optimization problem, and estimate the HR image, optical flow and blur kernel alternately or simultaneously. Since a large number of iterations are needed to reach convergence, the learning-based VSR methods are also time-consuming.

5.3 Deep Learning based VSR Method

As we analyzed above, the hand-crafted VSR approaches treat SR as a sophisticated optimization problem, which require expensive computational time and suffer considerable inference cost. Furthermore, the hand-crafted VSR approaches are not always applicable for practical scenarios where the imaging process may have different properties than assumed in the learning stage, leading to degraded performance (Yang et al. 2018).

Recently, deep learning based VSR methods have been proposed, with explicit or implicit temporal alignment. They have become the dominant technique for VSR (Jo et al. 2018; Lucas et al. 2019) since deep networks have strong model capacity to learn useful priors for VSR from a large video dataset, and can be trained end-to-end. Figure 11 shows the framework of the Deep Learning (+ optical flow) based VSR method.

According to how the temporal dependency among successive LR frames is exploited, the deep learning based VSR techniques have two main strategies (Liao et al. 2015; Kappeler et al. 2016; Liu et al. 2018): (a) utilizing convolutional neural networks (CNNs), and performing motion compensation explicitly to align LR video frames as the input for the CNN model (Liu et al. 2018), (b) using recurrent networks

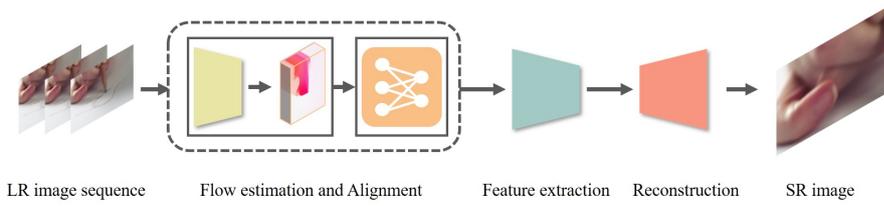


Fig. 11 The flow-chart of the Deep Learning (+ optical flow) based VSR method. Specifically, in the “Flow estimation and Alignment” module, optical flow estimation is conducted on LR video frames for motion compensation, and the alignment is achieved by learning a mapping from compensated LR video frames. In the “Feature extraction” module, the deep feature is extracted with a deep neural network (e.g. CNN, RNN). In the “Reconstruction” module, the HR video frames are super-resolved for the corresponding LR video frames.

(RNNs) to capture the temporal dependency (Haris et al. 2019) to avoid perform motion estimation explicitly.

5.3.1 CNN (+ Optical Flow) based VSR Method

The CNN based VSR methods are usually conducted in the following way: firstly, estimating optical flow from LR video frames for motion compensation; secondly, adjacent LR frames are motion compensated and utilized as the input to a CNN to produce HR images (Kappeler et al. 2016; Sajjadi et al. 2018; Wang et al. 2020b). We termed this kind of approaches as the CNN (+ Optical Flow) based VSR method, called CNN based VSR method for short.

Since Dong et al. 2014 proposed to use CNN to conduct single image SR and achieved the state-of-the-art performance, CNN is widely investigated for SR. Currently, the CNN based VSR method has become the dominant technique due to the following advantages:

- Once a CNN is trained, super-resolving an image is just a feed-forward process, making it is much more efficient than the traditional VSR methods;
- Neural networks have powerful learning capacity to model the spatial relation of the video frames, especially when with sufficient video data;
- The CNN framework usually can be trained end-to-end.

(A). Temporal Concatenation

To model temporal information in VSR, one of the most popular methods is to concatenate the frames (Liao et al. 2015; Kappeler et al. 2016; Caballero et al. 2017; Wang et al. 2020b). However, this strategy is unable to represent multiple motion regimes on a sequence as the input video frames are directly concatenated together (Haris et al. 2019). Furthermore, it is hard to train the network since many frames are processed simultaneously.

Deep-DE: Liao et al. 2015 proposed a deep draft-ensemble (Deep-DE) learning SR framework for fast VSR. They integrate SR drafts via the nonlinear process in a convolutional neural network (CNN) to recover high-frequency details. The SR

draft-ensemble process can generate several SR drafts according to a set of motion estimates, i.e., optical flows, where the optical flows are computed via different parameter settings. The architecture of the Deep-DE model is depicted in Figure 12.

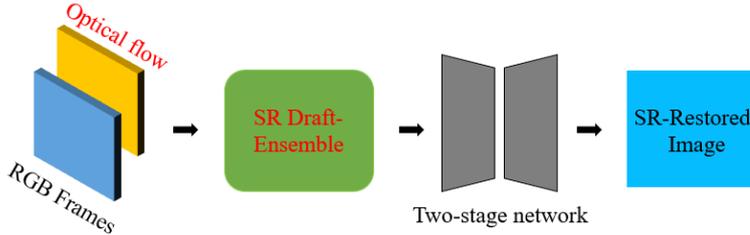


Fig. 12 The architecture of the Deep-DE model.

VSR-Net: Kappeler et al. 2016 presented to use CNN for VSR, where the CNN is trained on both the spatial and the temporal dimensions of videos to boost their spatial resolution. To obtain motion compensated frames as input for CNN, an adaptive motion compensation approach is introduced, in which the motions of input LR frames are compensated via a traditional optical flow algorithm. Then, the compensated frames are concatenated and fed to a pre-trained CNN SR network to reconstruct the SR frame. The optical flow algorithm (Drulea and Nedevschi 2011), which is a combination of the Local-Global approach with Total Variation (CLG-TV), is used. Besides, this adaptive motion compensation approach is able to address the issues of fast moving objects and motion blur in videos. The architecture of the VSR-Net model is depicted in Figure 13.

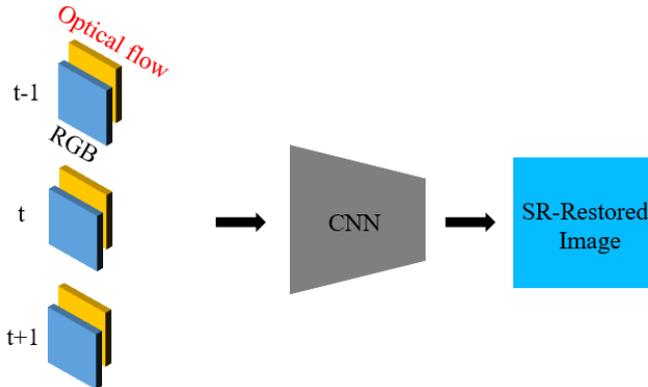


Fig. 13 The architecture of the VSR-Net model.

VESPCN: Methods Deep-DE (Liao et al. 2015) and VSR-Net (Kappeler et al. 2016) use a two-step framework, which separate the motion estimation from the net-

work training. Consequently, they are hard to get an overall optimal solution. To address this issue, Caballero et al. 2017 proposed a VESPCN method, which exploits a trainable motion compensation network, and utilizes a CNN to produce HR predictions from multiple LR frames in an end-to-end manner. It is the first end-to-end CNN based VSR framework. It takes advantage of sub-pixel convolution, temporal redundancy extraction via the spatio-temporal network, and the motion compensation, resulting in improved VSR performance in both accuracy and efficiency. Particularly, they designed an efficient spatial motion compensation transformer (MCT) module to estimate and compensate the motion between video frames in terms of optical flow, where the optical flow is computed in a coarse-to-fine strategy. After that the compensated frames are fed into the convolutional network for feature extraction and fusion. At last, the super-resolution process is conducted through a sub-pixel convolutional layer. The architecture of the VESPCN model is depicted in Figure 14.

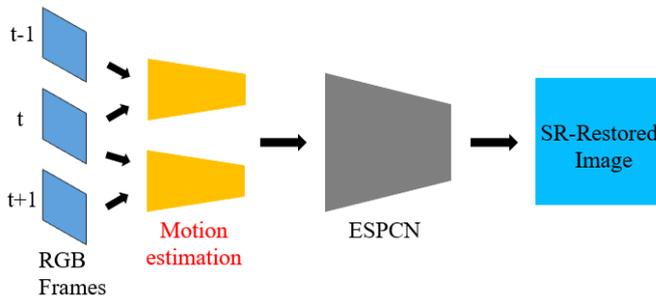


Fig. 14 The architecture of the VESPCN model.

STTN: Most optical flow techniques process only a pair of video frames (Kim et al. 2018), which are sensitive to complex situations like noise, occlusion and illumination change, and thus optical flow is limited for VSR. To overcome this disadvantage of the optical flow approaches, Kim et al. 2018 presented a spatio-temporal transformer network (STTN), which is able to handle multiple frames at a time. STTN consists of three main components: (a) a spatiotemporal flow estimation module, (b) a spatio-temporal sampler module, and (c) a super-resolution module. In particular, (a) the spatio-temporal flow estimation module is a U-Net style network (Ronneberger et al. 2015), which can estimate the optical flow of successive input frames including the target frame and multiple neighboring frames. The final result is a 3-channel spatio-temporal flow that describes the spatial and temporal changes between multiple video frames. (b) The spatio-temporal sampler module is in fact a trilinear interpolation approach, which is used to conduct warping for the current multiple neighboring frames and obtain the aligned video frames in terms of the spatio-temporal optical flow that is gained by the spatio-temporal flow module. (c) The super-resolution module is applied to perform feature fusion and super-resolution reconstruction for the target frame. The architecture of the STTN model is depicted in Figure 15.

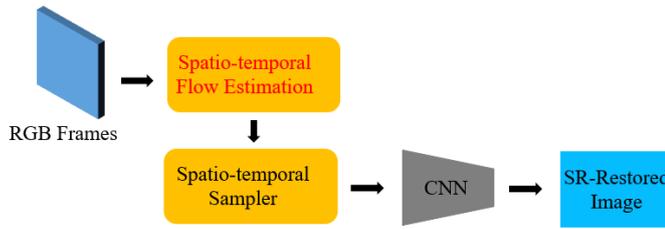


Fig. 15 The architecture of the STTN model.

TOFlow: Xue et al. 2019 designed a task-oriented flow (TOFlow) architecture, which combines the optical flow estimation network with the super-resolution reconstruction network, and trains these two networks jointly to compute optical flow. TOFlow adopts the framework of SpyNet (Ranjan and Black 2017) for optical flow estimation, and uses a spatial transformer approach to warp the neighboring frames based on the calculated optical flow. The final super-resolution is implemented by an image processing module. The architecture of the TOFlow model is depicted in Figure 16.

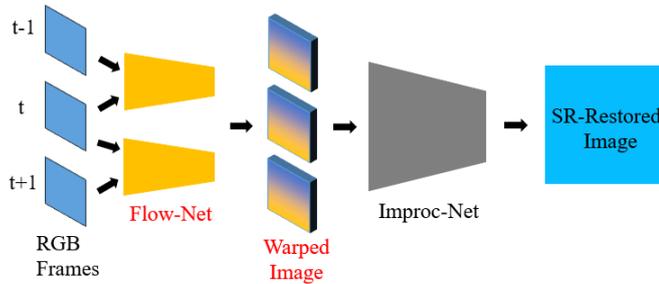


Fig. 16 The architecture of the TOFlow model.

MMCNN: To better extract spatiotemporal correlations between successive LR video frames and find more realistic details, Wang et al. 2019 presented a multi-memory CNN (MMCNN) for VSR, in which an optical flow network and an image-reconstruction network are cascaded. By embedding convolutional long short-term memory into the residual block, they designed a multi-memory residual block to replace the ordinary single-memory module to learn and retain inter-frame temporal correlations between the adjacent LR frames gradually. Specifically, the optical flow network can allow consecutive frames to serve as reference frames, thus it is beneficial for fusing multi-frame information. The architecture of the MMCNN model is depicted in Figure 17.

SoSR-ToSR: Zhang et al. 2019 applied VSR as a preprocessing step before feeding LR video frames into a two-stream action recognition network to address the issue that action recognition methods are un-applicable on low resolution videos. Specifically, to improve the performance of VSR, for the spatial stream, they designed an op-

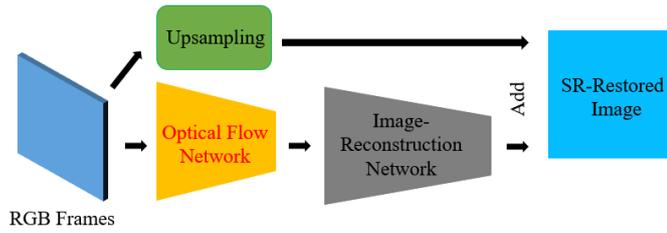


Fig. 17 The architecture of the MMCNN model.

tical flow guided weighted MSE loss to guide a spatial-oriented SR (SoSR) network to focus more on the regions with motion. For the temporal stream, they exploited a temporal-oriented SR (ToSR) network to enhance the adjacent frames together to ensure the temporal consistency. The architecture of the SoSR-ToSR model is depicted in Figure 18.

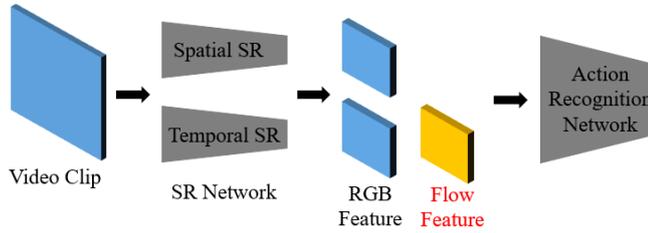


Fig. 18 The architecture of the SoSR-ToSR model.

MEMC-Net: Inspired by EDSR (Lim et al. 2017), Bao et al. 2019 proposed a motion estimation and motion compensation network (MEMC-Net) for VSR. The main contribution of MEMC-Net is the exploited adaptive warping layer, which warps the neighboring video frames by the computed optical flow and the convolutional kernel. The FlowNet (Dosovitskiy et al. 2015) is used for the motion estimation network, and a modified U-Net (Ronneberger et al. 2015), which consists of five max-pooling layers, five un-pooling layers and skip connections from the encoder to the decoder, is utilized for the kernel estimation network. To handle the occlusion problem, MEMC-Net extracts the feature of input frames by a pre-trained ResNet18, and feeds the output of the first convolutional layer of ResNet18 as the context information into the adaptive warping layer to conduct the warping again. the architecture of the MEMC-Net model is depicted in Figure 19.

MultiBoot-VSR: Kalarot et al. 2019 proposed a multi-stage multi-reference bootstrapping method for VSR (MultiBoot-VSR). MultiBoot-VSR is a two-stage framework, where the output of the first stage is utilized as the input of the second stage to further boost the performance. Initially, the FlowNet 2.0 (Ilg et al. 2017) algorithm is adopted to estimate optical flow and operate motion compensation. After that, the processed video frames are fed into the first-stage network to super-resolve the target

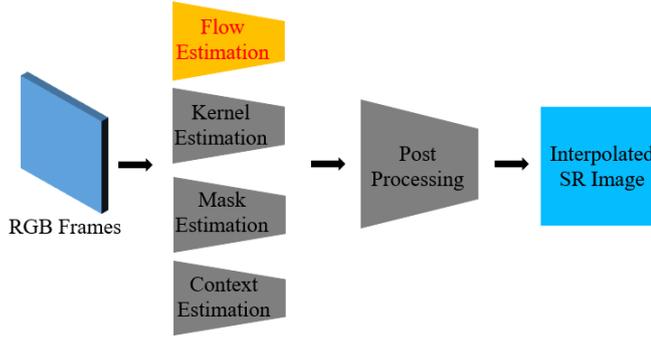


Fig. 19 The architecture of the MEMC-Net model.

frame. Lastly, the output from the first-stage is downsampled, concatenated with the original LR frame, and input to the second stage MultiBoot network to compute the final super-resolution result of the target frame. The architecture of the MultiBoot-VSR model is depicted in Figure 20.

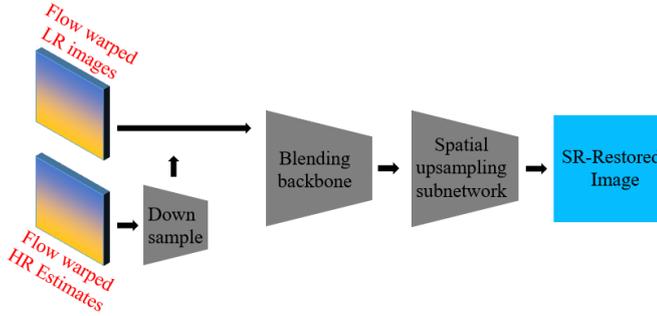


Fig. 20 The architecture of the MultiBoot-VSR model.

SOF-VSR: To handle the problem of resolution conflict between LR optical flows and HR outputs, where the resolution conflict prevents the recovery of fine temporal details, Wang et al. 2020b presented an end-to-end network to super-resolve optical flows for VSR, which is named SOF-VSR. In particular, this work contains three main steps: (a) an optical flow reconstruction network (OFR-Net) is used to infer HR optical flows in a coarse-to-fine way; (b) motion compensation is conducted via the estimated HR optical flows to encode temporal dependency; (c) the compensated LR images are input to a super-resolution network (SR-Net) to produce SR images. The architecture of the SOF-VSR model is depicted in Figure 21.

DDAN: Li et al. 2020 explored a deep dual attention network (DDAN), which consists of two main components, i.e., a motion compensation network (MCNet) and a SR reconstruction network (ReconNet), to fully exploit the spatio-temporal dependencies and learn discriminative spatio-temporal features for accurate VSR. To be

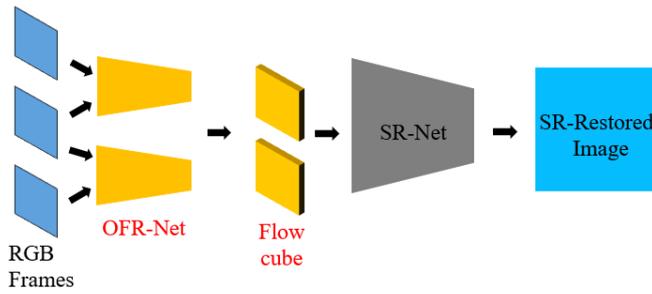


Fig. 21 The architecture of the SOF-VSR model.

specific, (1) the MCNet progressively learns multi-scale optical flow representations to synthesize the motion information across adjacent frames in a pyramid coarse-to-fine manner. To reduce the mis-registration errors caused by the optical flow based motion compensation, DDAN (Li et al. 2020) took two effective measures. First, except adopting the commonly used downscaling motion estimation strategy, it also utilizes a module without any downsampling operation to capture full resolution optical flow representations for better motion compensation. Second, the prior optical flow based methods usually simply concatenate the compensated frames and center frame for feature extraction and reconstruction, where the errors in the estimated optical flow or wrapping will adversely impact the subsequent SR reconstruction and bring artifacts. In contrast, DDAN extracts the detail components of original LR neighboring frames as complementary information to alleviate the errors of motion estimation. (2) In the ReconNet, DDAN incorporates the dual attention mechanism along channel and spatial dimensions with residual learning to focus on the intermediate informative features for high-frequency details recovery. The MCNet and ReconNet can be trained jointly in an end-to-end way for motion compensation and video SR reconstruction. The architecture of the DDAN model is depicted in Figure 22.

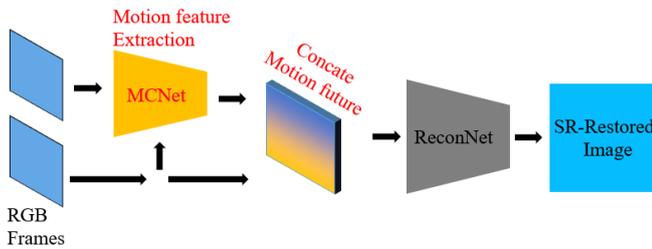


Fig. 22 The architecture of the DDAN model.

(B). Temporal Aggregation

To address the issue of dynamic motion in VSR, some methods presented multiple SR inferences to work on different motion regimes (Liu et al. 2018), where the outputs of all branches are aggregated at the last layer to construct the SR frame. However, these methods are difficult for global optimization because they also need to concatenate some input frames.

DRVSR: To address two important sub-problems in VSR, i.e., (a) aligning multiple frames to construct accurate correspondence and (b) fusing image details to produce high-quality results, Tao et al. 2017 proposed a detail-revealing deep video super-resolution (DRVSR) method, in which a “sub-pixel motion compensation” (SPMC) layer is exploited to conduct the up-sampling and motion compensation jointly for adjacent input frames according to the estimated optical flow. Specifically, DRVSR includes three main components, i.e., (a) a motion estimation module, (b) a motion compensation module, and (c) a fusion module. DRVSR respectively uses the motion compensation transformer (MCT) (Caballero et al. 2017) for motion estimation and the SPMC layer for motion compensation. Specifically, the SPMC layer applies sub-pixel information from the optical flow field to get sub-pixel motion compensation and resolution enhancement. For the fusion module, a detail fusion (DF) network is utilized to fuse image details from multiple video frames after SPMC alignment effectively. Additionally, the fusion module also utilizes a ConvLSTM module (Glorot and Bengio 2010) to tackle the spatio-temporal information. However, the SPMC layer costs large memory and has limited function, which prevents DRVSR to go deeper. The architecture of the DRVSR model is depicted in Figure 23.

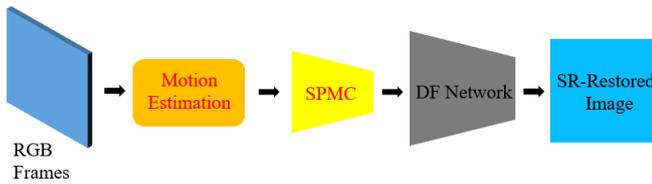


Fig. 23 The architecture of the DRVSR model.

TANN: Liu et al. 2017 presented a temporal adaptive neural network (TANN), which can tackle various types of movement robustly and choose the optimal range of temporal dependency automatically. In this way, useful information among successive video frames can be captured and the damage caused by erroneous motion can be reduced. They simplify the motion estimation in the patch level to incorporate translations to avoid interpolation. The rectified optical flow alignment is better than the traditional optical flow based image alignment in reconstructing the HR image. The architecture of the TANN model is depicted in Figure 24.

Limitation. The CNN based VSR methods generally suffer from the following limitations:

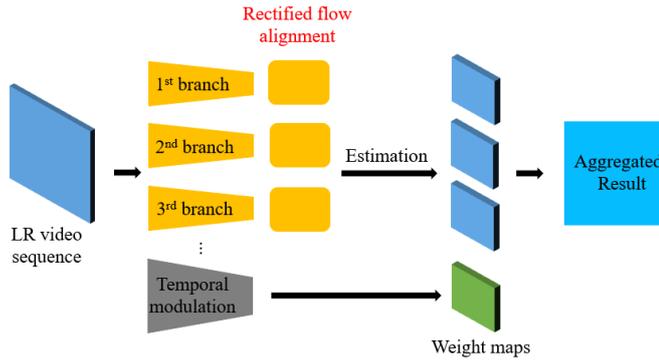


Fig. 24 The architecture of the TANN model.

- Dividing the VSR into a large number of separate multi-frame SR subtasks, resulting in temporal inconsistent, and unsatisfactory flickering artifacts would be produced.
- For the commonly used frame-concatenation strategy, many frames are processed simultaneously in the network, leading to hard training.
- Much time is wasted as each input frame is processed several times in the network.

5.3.2 RNN (+ Optical Flow) based VSR Method

Optical flow estimation usually costs high computational cost. To boost the efficiency, recurrent neural networks (RNNs) are now widely used for VSR, after the pioneering work of (Huang et al. 2015; Xiang et al. 2020). Generally, compared to CNN based VSR approaches, in the RNN based VSR framework, implicit temporal alignment is performed in terms of optical flow to replace explicit temporal alignment that depends on optical flow based motion estimation and compensation (Sajjadi et al. 2018; Yang et al. 2018). We termed this kind of approaches as RNN (+ Optical Flow) based VSR method, called RNN based VSR method for short.

FRVSR: Sajjadi et al. 2018 exploited an effective end-to-end trainable frame-recurrent VSR method (named FRVSR), which uses previously inferred HR estimates to super-resolve the subsequent video frames. Due to the application of the recurrent architecture, two benefits are gained: (a) Reducing the computational cost as each input frame is only processed once. (b) Enhancing the ability of the network to produce temporally consistent frames. Because the information from past frames will be passed to later frames through the HR estimates which are recurrently propagated over time. The FRVSR framework contains two important components, i.e., (a) the optical flow estimation network FlowNet, and (b) the super-resolution network SRNet. One distinct characteristic of FRVSR is its alignment strategy, where it does not warp the prior frame of the target directly while warps the HR version of the prior frame instead. However, since FRVSR simply refers to previously inferred HR frames, serious jitter and jagged artifacts are generated due to the former super-

resolving errors are accumulated to the subsequent frames. The architecture of the FRVSR model is depicted in Figure 25.

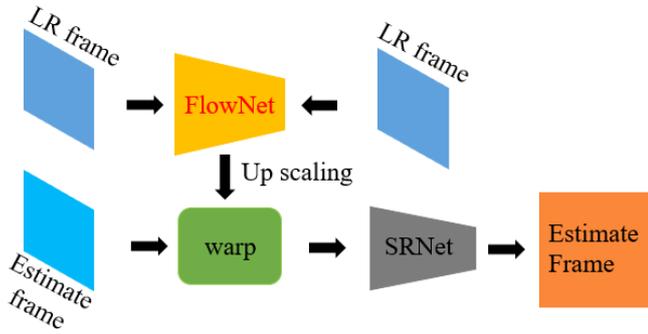


Fig. 25 The architecture of the FRVSR model.

STR-ResNet: Yang et al. 2018 exploited a Spatial Temporal Recurrent Residual Network (STR-ResNet) for video SR, which is able to model intra-frame redundancy and inter-frame motion context jointly in a unified deep framework, due to the framework combines the spatial convolutional and temporal recurrent architectures. This network does not require explicit optical flow estimation for motion compensation. The architecture of the STR-ResNet model is depicted in Figure 26.

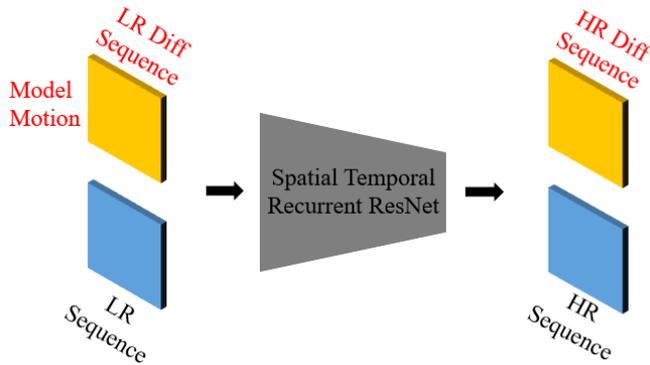


Fig. 26 The architecture of the STR-ResNet model.

RRCN: Li et al. 2019 presented a very deep non-simultaneous fully recurrent convolutional network for VSR. Specifically, they applied very deep fully recurrent convolutional layers and late fusion on motion compensated video frames to make full use of the temporal information in their non-simultaneous recurrent convolutional network architecture. A CLG-TV optical flow method (Drulea and Nedevschi 2011) is utilized for motion estimation, and the centering frame is chosen as the reference frame to compensate the adjacent frames, then both the centering frame and the

motion compensated frames are fed into their network for VSR. Remarkably, the very deep recurrent convolutional network has powerful representation ability. Besides, it is good at modeling the spatial and temporal non-linear mappings. The architecture of the RRCN model is depicted in Figure 27.

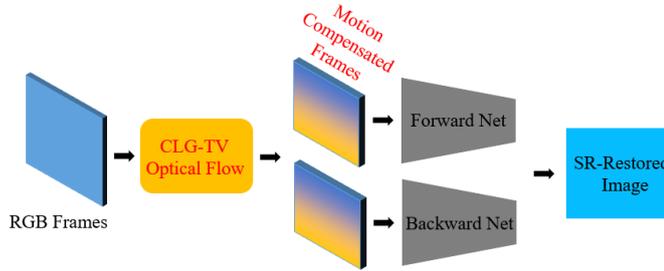


Fig. 27 The architecture of the RRCN model.

RBP: Haris et al. 2019 proposed a Recurrent Back-Projection Network (RBP), which consists of three main components: a feature extraction module, a projection module, and a reconstruction module. RBP collects the spatial and temporal information from video frames surrounding the target one. Back-projection in the recurrent process is used to organize the temporal information, where high-resolution features can be gradually refined and applied to reconstruct the high resolution target frame. Optical flow is used for initial feature extraction. Specifically, before entering projection modules, they contact the pre-computed optical flow motion maps with the target frame I_t in the corresponding neighborhood $[I_{t-N}, \dots, I_{t-k}, \dots, I_t]$, N is the temporal neighborhood frame numbers. The optical flow motion map can encourage the projection module to capture missing details between I_t and its neighbors I_{t-k} . For the reconstruction part, DBPN (Haris et al. 2018) is employed as the single image super-resolution network, and ResNet (He et al. 2016) with deconvolution is utilized as the multi-image super-resolution network. The architecture of the RBP model is depicted in Figure 28.

Limitation. RNN based VSR methods generally suffer from the following limitations:

- It is good at modelling global slow-varying motions but not those short-term fast-varying ones. This is because the recurrent connections operate on hidden states, while significant fine-grained details for depicting fast-varying motions mostly exist in input video frames other than the hidden states (Baker and Kanade 1999; Huang et al. 2018).
- Without explicit temporal alignment, the RNN based VSR methods have limited ability to deal with complex and large motions.
- The structure information, which is important for super-resolving LR video frames, is lost. This is caused by the dimensionality reduction when transforming the input 2D LR video frames to the 1D vectors of RNN hidden states.

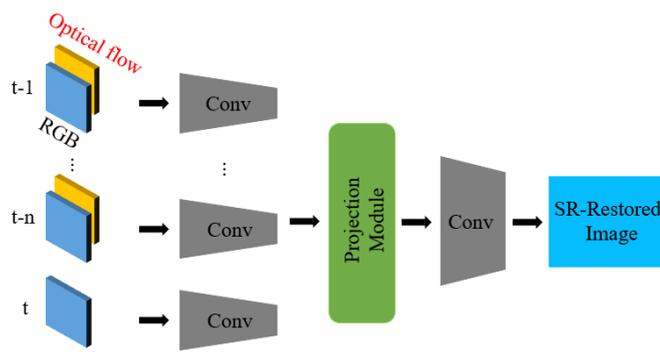


Fig. 28 The architecture of the RBP model.

Table 1 Comparison of CNN (+ Optical Flow) based VSR Methods. MEC denotes Motion estimation & compensation.

Method	Year	MEC	Advantage
Deep-DE	2015	Explicit MEC, Optical flow	<ul style="list-style-type: none"> 1) SR draft ensemble generation: Renovate traditional feed-forward reconstruction pipeline; Enhance ability to compute different super-resolution results; Consider large motion variation and various latent artifacts. 2) SR draft based optimal reconstruction: Using CNN to integrate the reconstruction and deconvolution steps; Avoiding parameter tuning in the test phase.
VSR-Net	2016	Explicit MEC, Optical flow	<ul style="list-style-type: none"> 1) Proposed a CNN trained on both the spatial and temporal spaces to enhance the spatial resolution; 2) Integrating motion compensated frames, filter symmetry enforcement, and pre-training strategy to improve both the accuracy and efficiency; 3) Exploited an adaptive motion compensation method to handle motion blur and fast moving objects.
VESPCN	2017	Explicit MEC, Optical flow	<ul style="list-style-type: none"> 1) Studying different fusion approaches to discover spatio-temporal correlations; 2) Building a motion compensation scheme based on a fast multi-resolution spatial transformer module; 3) Exploiting a spatio-temporal sub-pixel convolution network to improve accuracy and temporal consistency of VSR in real-time.
STTN	2018	Explicit MEC, Optical flow	<ul style="list-style-type: none"> 1) Exploring a spatio-temporal flow estimation network to selectively capture long-range temporal dependency efficiently and handle occlusion effectively; 2) Presenting a spatio-temporal sampler to enable spatio-temporal manipulation; 3) Integrating on the top of conventional networks easily.
TOFlow	2019	Explicit MEC, Optical flow	<ul style="list-style-type: none"> 1) Proposed a TOFlow network to learn flow representation in a self-supervised, task-specific way; 2) Jointly learn the task-oriented optical flow and perform VSR; 3) The network is end-to-end trainable.
MMCNN	2019	Explicit MEC, Optical flow	<ul style="list-style-type: none"> 1) Designing a MMCNN for accurate and fast VSR by cascading an optical flow module and an image-reconstruction module; 2) Replacing sub-pixel motion compensation with motion transformer operator to faster optical flow estimation; 3) Embedding LSTM into the residual block to form a multi-memory residual block to progressively learn and retain temporal dependency between adjacent LR frames; 4) Using a series of residual blocks engaged in terms of intra-frame spatial correlations for feature extraction and reconstruction.

Table 1 (Continued).

Method	Year	MEC	Advantage
SoSR-ToSR	2019	Explicit MEC, Optical flow	<ol style="list-style-type: none"> 1) Designing a two-stream network for VSR; 2) Proposing SoSR with optical flow guided weighted MSE loss to pay more attention to moving objects; 3) Exploiting ToSR with a siamese network to emphasize temporal consistency.
MEMC-Net	2019	Explicit MEC, Optical flow	<ol style="list-style-type: none"> 1) Proposed a motion estimation and compensation driven neural network for robust and high-quality video frame interpolation; 2) Exploited an adaptive warping layer to integrate both optical flow and interpolation kernels to synthesize target frame pixels; 3) This adaptive warping layer is fully differentiable to enable to jointly optimize the flow and kernel estimation networks.
MultiBoot-VSR	2019	Explicit MEC, Optical flow	<ol style="list-style-type: none"> 1) Proposed a scene and class agnostic, fully convolutional neural network; 2) The model consists of a motion compensation based input subnetwork, a blending backbone, and a spatial upsampling subnetwork; 3) Reusing reconstructed high-resolution frames as additional reference frames after reshuffling them into multiple low-resolution images to bootstrap and enhance image quality progressively.
SOF-VSR	2020	Explicit MEC, Optical flow	<ol style="list-style-type: none"> 1) Designed a unified SOF-VSR to jointly super-resolve optical flow and images; 2) Exploited an OFRnet to infer HR optical flows from LR frames in a coarse-to-fine manner to recover accurate temporal details for SR; 3) Motion compensation is conducted via HR flows to encode temporal dependency.
DRVSR	2017	Explicit MEC, Optical flow	<ol style="list-style-type: none"> 1) Can take arbitrary-size input images; 2) Exploited a SPMC to better handle inter-frame motion; 3) SPMC layer can be used for arbitrary scaling factors during testing as it without trainable parameters; 4) A DF network is applied to effectively fuse image details from multiple images after SPMC alignment.
TANN	2018	Explicit MEC, Optical flow	<ol style="list-style-type: none"> 1) Proposed a DDAN for VSR which includes a MCNNet and a ReconNet to fully exploit the spatio-temporal informative features; 2) MCNNet progressively extracts flow representations to synthesize the motion information across adjacent frames in a pyramid fashion; 3) Capturing detail components of original LR adjacent frames as complementary cues for accurate feature extraction to reduce mis-registration errors of motion estimation; 4) Conducting dual attention on a residual unit and form a residual attention unit to emphasize meaningful features for high-frequency details recovery.

Table 1 (Continued).

Method	Year	MEC	Advantage
STR-ResNet	2018	Implicit MEC, RNN	<ol style="list-style-type: none"> 1) Explored a STR-ResNet VSR method by jointly modeling intra-frame redundancy and inter-frame motion context in a unified deep neural network; 2) The STR-ResNet is the first study attempts to incorporate the bypass connection in a deep network to embed the joint spatial-temporal residue prediction and model temporal correlations in video frames; 3) The network is able to implicitly model the motion context among multiple video frames for VSR; 4) It among the first to investigate and integrate the spatial convolutional, temporal recurrent and residual architectures into a single deep network to address VSR.
FRVSR	2018	Implicit MEC, RNN	<ol style="list-style-type: none"> 1) Exploited an end-to-end trainable frame-recurrent VSR method by using the previously inferred HR estimate to super-resolve the subsequent video frame; 2) Ensuring temporally consistent and reducing the computational cost by warping only one image in each step; 3) Enabling to assimilate a large number of prior frames without increased computational requirement; 4) Without needing any pre-training stages.
RBPV	2019	Implicit MEC, RNN	<ol style="list-style-type: none"> 1) A RBPV is proposed to treat each frame as a separate source, the sources are integrated in an iterative refinement framework via back-projection modules; 2) Integrating SISR and MISR in a unified VSR framework; 3) Explicitly representing estimated inter-frame motion with respect to the target instead of explicitly aligning frames.
RRCN	2019	Explicit MEC, Optical flow, RNN (Model temporal dependency)	<ol style="list-style-type: none"> 1) Proposed the first very deep non-simultaneous fully RNN for VSR; 2) A model ensemble method is exploited to combine multi-frame SR model with single-image SR model; 3) Strong representation ability; 4) Good at modeling the spatial and temporal non-linear mappings.

6 Challenge and Future Direction

Although great progress has been made for VSR in the past decades, there remain some open research questions. In this section, we will discuss these challenges explicitly and outline some promising directions for future study. The challenges and trends are investigated in two aspects.

6.1 Challenges for optical flow-based temporal alignment

Since accurate optical flow estimation remains challenging for real videos, the optical flow-based temporal alignment is still a key problem in VSR, where the main challenges including:

- Inaccurate flow will result in distortion and errors, deteriorating the final VSR performance.
- Image-level warping strategy introduces artifacts into the aligned frames.
- Fast moving and large displacement are difficult to handle, which significantly affect the performance of both optical flow estimation and flow warping.
- Per-pixel motion estimation suffers a heavy computational cost.

6.2 Future direction for video super-resolution

6.2.1 *Lightweight VSR Model*

Currently, most VSR methods emphasize on pursuing high performance, leading to large models (contain a huge number of parameters) that take a long time for training while require high computing and storage resources. These characteristics prevent their use on mobile devices in practical applications, where small models and fast inference speed are preferred. Therefore, how to design lightweight VSR models while still maintaining high performance is a promising research topic.

6.2.2 *Applicable to Real-world Scenarios*

VSR methods face difficulties in real-world scenarios as they often suffer from issues like unknown degradation, scene change, occlusion, etc. Boosting the ability of VSR methods to handle such real-world scenarios is urgent.

- **Handling Degradation.** Existing VSR methods normally generate LR video frames according to the manners of downsampling directly with interpolation (e.g. bicubic interpolation) or downsampling after Gaussian blurring, manually. However, it is well-known that the real-world degradation process is very complicated which includes many uncertainties. As a result, VSR models, which are trained on artificially produced degradation by blurring and interpolation, cannot well conformed to actual LR video frames in practice. Consequently, better ways of modeling and handling degradation are needed.

- **Handling complex scene changes.** In reality, a video often contains some different scenes. However, the current VSR methods cannot deal with such changes well. A typical approach is split the videos into multiple segments, each without scene changes, and then process them individually. This kind of strategy increases the computational cost, therefore new ways to handle videos with scene changes are necessary for realistic VSR applications.

6.2.3 Effectively Exploring Spatio-temporal Information

Except the appearance RGB information, video includes the temporal information. How to effectively explore temporal information across video frames will directly affect the performance of VSR. Current methods, like 3D convolution and non-local modules are computationally inefficient, and the quality of optical flow cannot be guaranteed. Consequently, proposing new methods to effectively make use of spatio-temporal information in video is worth further study.

6.2.4 More Reasonable Evaluation Metrics

Evaluation metric is the most fundamental component in each computer vision task, which greatly influences the task's progress. Exploring more reasonable evaluation metrics is of equal importance to exploit more advanced algorithms.

Similar to image SR, nowadays, the performance of video SR is also evaluated mainly by the FR-IQA metrics like peak signal-noise ratio (PSNR) and structural similarity index (SSIM) (Wang et al. 2004; Yang et al. 2019). These FR-IQA measures face some fatal challenges: (1) They are unable to reflect the video quality for human perception quite well. Because these methods depend on the pixel-level error measures, like L1 and L2 distances or their combination (Timofte et al. 2018), causing them concentrate on local pixel-level information, thus they cannot measure perceptual quality accurately (Ledig et al. 2017b; Ma et al. 2017). (Blau et al. 2018) has demonstrated that images with high PSNR and SSIM produce overly smooth images with low perceptual quality; (2) They are designed in a limited and refined condition, which require manual intervention, thus they lack of feasibility for modeling unknown distortions (Anwar et al. 2020). (3) They require non-distorted ground-truth images for comparison, which are almost unavailable in practice, because it is difficult or impossible to acquire ideal reference images in most conditions, especially for the real video data.

To address this issue, some new perceptual-based NR-IQA metrics have been proposed (Kim and Lee 2017; Talebi and Milanfar 2018; Zhang et al. 2018b; Prashnani et al. 2018; Tang et al. 2019; Zhu et al. 2020). However, there are still no universally accepted evaluation criteria that can work in various situations and perfectly assess SR quality. Even worse, we do not make clear what kind of perceptual quality is real important and useful for assessing SR currently. Nevertheless, proposing new effective metrics which can be broadly used is remaining an open research problem.

6.2.5 Unsupervised VSR

The state-of-the-art VSR approaches are deep neural network based and trained in the supervised manner. However, deep learning networks require a large number LR-HR video frame pairs for training. On the one hand, the paired datasets are rare or costly to acquire in practice. On the other hand, when the input video frames are poor in resolution, the super-resolution cannot work well. Furthermore, the current VSR models trained on these artificial labeled datasets can only learn the inverse process of the predefined degradation, which is too simple to characterize the real-world situation. One promising direction is to exploit unsupervised VSR methods which can be well performed on unpaired LR-HR video sets.

7 Conclusion

Video super-resolution (VSR), which aims to improve the clarity and visual appearance of video frames, is a crucial task in computer vision and has been deeply investigated. One core problem for VSR is how to capture temporal dependency to achieve efficient and accurate temporal alignment. Noticeably, the most popular way is to use optical flow to acquire motion information for explicit or implicit temporal alignment. In this work, we provide a comprehensive review of optical flow based VSR methods by introducing previous work and analyzing current advances, and discussing their limitations incidentally. In particular, we firstly explain what is video super-resolution. Secondly, we give a detailed explanation about what is optical flow based VSR. Thirdly, both the representative traditional (i.e. reconstruction based VSR method and learning based VSR method) and the current deep learning based VSR algorithms that make use of optical flow are compared and explored. Remarkably, we deeply investigate the deep neural network related VSR methods. Fourthly, we find that although deep learning based VSR algorithms have achieved great progress, there are still some practical issues and unsolved problems. Accordingly, we discuss the challenges and point out the promising future research trend for VSR. To the best of our knowledge, this is the first systematical work on surveying the effect of optical flow in VSR. We hope this survey would not only provide a better and deeper understanding of optical flow based VSR, but also serve as a catalyst to spur future research activities in this domain.

To promote the future research activity, we hope to play as a forerunner and start the discussion from the following aspects:

- 1) **Lightening VSR Model with Knowledge Distillation.** There is a considerable performance gap between the lightweight VSR model and the normally used complex VSR model, while the latter one requires a much larger amount of resources (Xiao et al. 2021). This problem is particularly acute on resource-limited devices, e.g., smartphones and wearable devices. Where a compact VSR model can be easily used on these devices, but due to its limited capacity to model spatial-temporal correlations, the VSR performance is unsatisfactory. Knowledge distillation, which is able to transfer knowledge from a complicated model (teacher network) to a simplified one (student network), and without altering the original architecture of the teacher net-

work, supplies a possible way to handle this problem. Designing a spatial-temporal distillation (STD) scheme, which is suitable for VSR, is a promising research topic.

2) Promoting VSR for Real-world Scenarios. Although VSR methods have achieved remarkable progress recently, they are still unapplicable in reality. One reason is that most of the existed VSR models are trained and assessed on synthetic datasets, where the videos are generated by simple synthetic degradation methods. Badly, these degradation methods unable to well simulate the complicated degradation processes in realistic videos, and accordingly leading to the trained VSR models noneffective for real-world applications (Yang et al. 2021). Research Topic 1: building a real-world video super-resolution dataset, which can bridge the synthetic-to-real gap in VSR and supply a valid benchmark for training and evaluating the real-world VSR models generally. The second reason is that the prior degradation models take some factors into account, e.g. blur, downsampling, noise, but they are still cannot cover the diverse degradations of real video data (Pan et al. 2021). Research Topic 2: exploiting a practical degradation model that consists of random degradations. For example, designing a VSR which can simultaneously estimate unknown blur kernels, motion fields, and latent HR videos effectively is prospective. Lots of VSR models have been proposed, but there is not a unified framework being dominant for VSR in practice yet (Chan et al. 2021; Yi et al. 2021). Research Topic 3: Exploring a generic, efficient, and easy-to-implement baseline framework for VSR, which can serve as a standard for various comparison and evaluation.

3) Exploiting More Useful Spatio-temporal Information. Capturing spatio-temporal information accurately and efficiently is critical important for VSR. Recently, using the deformable convolution backbone to conduct spatio-temporal VSR on the feature space directly is popular as this strategy is fast (Xiang et al. 2020). However, these VSR models would only produce pre-defined intermediate frames, causing them constrained to highly-controlled scenarios with fixed frame-rate videos. Consequently, exploiting controllable spatio-temporal VSR approaches, which with the deformable convolution network, for smooth motion synthesizing it is necessary (Xu et al. 2021).

In addition, current VSR models underrated the short-term motion cues between successive video frames, therefore how to exploit both the short-term and long-term motion cues in videos is desirable.

4) Designing Natural Evaluation Metrics. Since the ground truth reference image is almost absent for real data, it means that the widely used FR-IQA metrics are unreasonable in fact, human evaluation is the only rational way to evaluate the performance of VSR models. Mean-Opinion-Score (MOS), i.e. the average rating that human raters assigned to super-resolved images via a certain SR model, is the typical human evaluation measure. However, the MOS values of different models are not directly comparable due to the changing of rater numbers and rater's subjectivity, etc (Khrulkov and Babenko 2021). Consequently, it is urgent to propose new NR-IQA metrics, which enable to break through the current evaluation predicament in some aspects: (1) comparing with various VSR models automatically, (2) approximating human preferences naturally; (3) tuning hyperparameters without artificial assistance effectively, etc.

5) Improving Unsupervised VSR. For unsupervised VSR, there are main two ways. First, super-resolving video images without introducing predefined degrada-

tion by using unpaired LR-HR datasets. Despite great progress, it is hard to synthesize good “real” LR images for super-resolution yet (Wang et al. 2021b). Formulating unpaired SR training as a domain adaptation issue, and enabling the VSR network to match LR images from different domains into a shared degradation-imperceptible feature space deserved to be discussed. Second, learning model with unsupervised deep networks, where Generative Adversarial Network (GAN) is the currently primary method. However, GAN usually brings noise and causes some details of dislocation. Exploring more advanced unsupervised deep networks is another research topic.

Acknowledgements This work was supported by the National Natural Science Foundation of China under Grant 62106177. It was also supported by the Central University Basic Research Fund of China (No.2042020KF0016, CCNU20TS028), the Teaching research project of CCNU (202013), and the Wuhan University-Infinova project No.2019010019. The numerical calculation was supported by the supercomputing system in the Super-computing Center of Wuhan University.

Conflict of interest

The authors declare that they have no conflict of interest.

References

1. Park, S., Park, M., Kang, M. (2003). Super-resolution image reconstruction: A technical overview. *IEEE Signal Processing Magazine*, 20(3), 21–35.
2. Fookes, C., Lin, F., Chandran, V., Sridharan, S. (2004). Super-Resolved Face Images using Robust Optical Flow. *IEEE Workshop on the Internet, Telecommunications and Signal Processing*, pp.391–396.
3. Nguyen, K., Fookes, C., Sridharan, S., Tistarelli, M., Nixon, M. (2018). Super-resolution for biometrics: A comprehensive survey. *Pattern Recognition*, 78:23–42.
4. Xiong, Z., Sun, X., Wu, F. (2010). Robust Web Image/Video Super-Resolution. *IEEE Transactions on Image Processing*, 19(8), 2017–2028.
5. Singh, A., Singh, J. (2020). Survey on single image based super-resolution-implementation challenges and solutions. *Multimedia Tools and Applications*, 79(3), 1641–1672.
6. Ledig, C., Theis, L., Huszar, F., Caballero, J., Cunningham, A., Acosta, A., Aitken, A., Tejani, A., Totz, J., Wang, Z., Shi, W. (2017a). Photo-Realistic Single Image Super-Resolution Using a Generative Adversarial Network. *In Proc. CVPR*, pp.4681–4690.
7. Ma, C., Yang, C., Yang, X., Yang, M. (2017). Learning a no-reference quality metric for single-image super-resolution. *Computer Vision and Image Understanding*, 158:1–16.
8. Farsiu, S., Robinson, M. D., Elad, M., Milanfar, P. (2004). Fast and Robust Multiframe Super Resolution. *IEEE Transactions on Image Processing*, 13(10), 1327–1344.
9. Shi, W., Caballero, J., Huszar, F., Totz, J., Aitken, A. P., Bishop, R., Rueckert, D., Wang, Z. (2016). Real-Time Single Image and Video Super-Resolution Using an Efficient Sub-Pixel Convolutional Neural Network. *In Proc. CVPR*, pp.1874–1883.
10. Talebi, H., Milanfar, P. (2018). NIMA: Neural Image Assessment. *IEEE Transactions on Image Processing*, 27(8), 3998–4011.
11. Hou, L., Yu, C., Samaras, D. (2016). Squared Earth Mover’s Distance-based Loss for Training Deep Neural Networks. *arXiv:1611.05916*, 1–9.
12. Kim, J., Lee, S. (2017). Deep Learning of Human Visual Sensitivity in Image Quality Assessment Framework. *In Proc. CVPR*, pp.1969–1977.
13. Nasrollahi, K., Moeslund, T. B., (2014). Super-resolution: a comprehensive survey. *Machine Vision and Applications*, 25:1423–1468.
14. Thapa, D., Raahemifar, K., Bobier, W., Lakshminarayanan, V. (2016). A performance comparison among different super-resolution techniques. *Computers and Electrical Engineering*, 54:313–329.

15. Wang, Z., Chen, J., Hoi, S. (2021a). Deep Learning for Image Super-resolution: A Survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(10), 3365–3387.
16. Tsai, R., Huang, T. S. (1984). Multiframe image restoration and registration. *Advances in Computer Vision and Image Processing*, pp.317–339.
17. Gerchberg, R. (1974). Super-resolution through error energy reduction. *Journal of Modern Optics*, 21(9), 709–720.
18. Mjolsness, E. (1985). Neural networks, pattern recognition, and fingerprint hallucination. *Ph.D thesis, California Institute of Technology, US*.
19. Peleg, S., Keren, D., Schweitzer, L. (1987). Improving image resolution using subpixel motion. *Pattern Recognition Letter*, 5(3), 223–226.
20. Cheeseman, P., Kanefsky, B., Kraft, R., Stutz, J. (1994). Super-resolved surface reconstruction from multiple images. *Technical Report FIA9412, NASA*.
21. Baker, S., Kanade, T. (2000). Hallucinating faces. *In Proc. Automatic Face and Gesture Recognition*, pp.83–88.
22. Shechtman, E., Caspi, Y., Irani, M. (2005). Space-time super-resolution. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(4), 531–45.
23. Yang, J., Wright, J., Huang, T., Ma, Y. (2010). Image super-resolution via sparse representation. *IEEE Transactions on Image Processing*, 19(11), 2861–2873.
24. Dong, C., Chen, C. L., He, K., and Tang, X. (2014). Learning a deep convolutional network for image super-resolution. *In Proc. ECCV*, pp.184–199.
25. Huang, Y., Wang, W., Wang, L. (2015). Bidirectional recurrent convolutional networks for multi-frame super-resolution. *In Proc. NIPS*, pp.235–243.
26. Liao, R., Tao, X., Li, R., Ma, Z., Jia, J. (2015). Video super-resolution via deep draft-ensemble learning. *In Proc. ICCV*, pp.531–539.
27. 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 Proc. CVPR*, pp.4778–4787.
28. Wang, X., Yu, K., Dong, C., Loy, C. (2018). Recovering Realistic Texture in Image Super-Resolution by Deep Spatial Feature Transform. *In Proc. CVPR*, pp.606–615.
29. Shocher, A., Cohen, N., Irani, M. (2018). Zero-shot super-resolution using deep internal learning. *In Proc. CVPR*, pp.3118–3126.
30. Yuan, Y., Liu, S., Zhang, J., Zhang, Y., Dong, C., Lin, L. (2018). Unsupervised image super-resolution using cycle-in-cycle generative adversarial networks. *In Proc. CVPRW*, pp.814–823.
31. Li, D., Liu, Y., Wang, Z. (2019). Video Super-Resolution Using Non-Simultaneous Fully Recurrent Convolutional Network. *IEEE Transactions on Image Processing*, 28(3), 1342–1355.
32. Zhai, G., Min, X. (2019). Perceptual image quality assessment: a survey. *Science China Information Sciences*, vol.63, pp.211301:1-52, 2020.
33. Shocher, A., Cohen, N., Irani, M. (2018). Zero-shot super-resolution using deep internal learning. *In Proc. CVPR*, pp.3118–3126.
34. Lin, K., Wang, G. (2018). Hallucinated-IQA: no-reference image quality assessment via adversarial learning. *In Proc. CVPR*, pp.732–741.
35. Zhu, H., Li, L., Wu, J., Dong, W., Shi, G. (2020). MetaQA: Deep Meta-learning for No-Reference Image Quality Assessment. *In Proc. CVPR*, pp.14143–14152.
36. Yuan, Q., Zhang, L., Shen, H., Li, P. (2010). Adaptive Multiple-Frame Image Super-Resolution Based on U-Curve. *IEEE Transactions on Image Processing*, 19(12), 3157–3170.
37. Borsoi, R. A., Costa, G. H., Bermudez, J. C. M. (2019). A New Adaptive Video Super-Resolution Algorithm With Improved Robustness to Innovations. *IEEE Transactions on Image Processing*, 28(2), 673–686.
38. Shen, C. T., Liu, H. H., Yang, M. H., Hung, Y. P., Pei, S. C. (2015). Viewing-Distance Aware Super-Resolution for High-Definition Display. *IEEE Transactions on Image Processing*, 24(1), 403–418.
39. Kappeler, A., Yoo, S., Dai, Q., Katsaggelos, A. (2016). Video Super-Resolution with Convolutional Neural Networks. *IEEE Transactions on Computational Imaging*, 2(2), 109–122.
40. Liu, C., Sun, D. (2014). On Bayesian adaptive video super resolution. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(2), 346–360.
41. Liu, D., Wang, Z., Fan, Y., Liu, X., Wang, Z., Chang, S., Huang, T. (2017). Robust Video Super-Resolution with Learned Temporal Dynamics. *In Proc. ICCV*, pp.2507–2515.
42. Mudunuri, S., Biswas, S. (2016). Low resolution face recognition across variations in pose and illumination. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(5), 1034–1040.

43. Chen, Y., Tai, Y., Liu, X., Shen, C., Yang, J. (2018). FSRNet: End-to-End Learning Face Super-Resolution with Facial Priors. *In Proc. CVPR*, pp.2492–2501.
44. Ma, C., Jiang, Z., Rao, Y., Lu, J., Zhou, J. (2020). Deep Face Super-Resolution With Iterative Collaboration Between Attentive Recovery and Landmark Estimation. *In Proc. CVPR*, pp.5569–5578.
45. Zhang, H., Liu, D., Xiong, Z. (2019). Two-Stream Action Recognition-Oriented Video Super-Resolution. *In Proc. ICCV*, pp.8799–8808.
46. Hong, C., Yu, J., Wan, J., Tao, D., Wang, M. (2015). Multimodal deep autoencoder for human pose recovery. *IEEE Transactions on Image Processing*, 24(12), 5659–5670.
47. Hore, A., Ziou, D. (2010). Image quality metrics: PSNR vs. SSIM. *In Proc. International Conference on Pattern Recognition*, pp.2358–2369.
48. Kim, K., Kwon, Y. (2010). Single-Image Super-Resolution Using Sparse Regression and Natural Image Prior. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(6), 1127–1133.
49. Sun, K., Xiao, B., Liu, D., Wang, J. (2019). Deep High-Resolution Representation Learning for Human Pose Estimation. *In Proc. CVPR*, pp.5693–5703.
50. Ryoo, M. S., Rothrock, B., Fleming, C., Yang, H. J. (2017). Privacy-Preserving Human Activity Recognition from Extreme Low Resolution. *In Proc. AAAI*, pp.4255–4262.
51. Jing, X., Zhu, X., Wu, F., Hu, R., You, X., Wang, Y., Feng, H., Yang, J. (2017). On Bayesian adaptive video super resolution. *IEEE Transactions on Image Processing*, 26(3), 1363–1378.
52. Baker, S., Schar, D., Lewis, J., Roth, S., Black, M., Szeliski, R. (2011). A database and evaluation methodology for optical flow. *International Journal of Computer Vision*, 92(1), 1–31.
53. Cruz, C., Mehta, R., Katkovnik, V., Egiazarian, K. O. (2018). Single Image Super-Resolution Based on Wiener Filter in Similarity Domain. *IEEE Transactions on Image Processing*, 27(3), 1376–1389.
54. Huang, Y., Lu, Z., Shao, Z., Ran, M., Zhou, J., Fang, L., Zhang, Y. (2019). Simultaneous denoising and super-resolution of optical coherence tomography images based on generative adversarial network. *Optics Express*, 27(9), 12289–12307.
55. Meur, O. L., Ebdelli, M., Guillemot, C. (2013). Hierarchical Super-Resolution-Based Inpainting. *IEEE Transactions on Image Processing*, 22(10):3779–3790.
56. Tan, W., Yan, B., Bare, B. (2018). Feature super-resolution: Make machine see more clearly. *In Proc. CVPR*, pp.3994–4002.
57. Cai, D., Chen, K., Qian, Y., Kamarainen, J. (2019). Convolutional low-resolution fine-grained classification. *Pattern Recognition Letters*, 119:166–171.
58. Zhao, H., Qi, X., Shen, X., Shi, J., Jia, J. (2018). ICNet for Real-Time Semantic Segmentation on High-Resolution Images. *In Proc. ECCV*, pp.418–434.
59. Wang, L., Li, D., Zhu, Y., Tian, L., Shan, Y. (2020a). Dual Super-Resolution Learning for Semantic Segmentation. *In Proc. CVPR*, pp.3774–3783.
60. Girshick, R., Donahue, J., Darrell, T., Malik, J. (2016). Region-based convolutional networks for accurate object detection and segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(1), 142–158.
61. Na, B., Fox, G. C. (2017). Object Detection by a Super-Resolution Method and a Convolutional Neural Networks. *In Proc. IEEE Int. Conf. Big Data*, pp.2263–2269.
62. Eekeren, A. W. M., Schutte, K., Vliet, L. J. (2010). Multiframe Super-Resolution Reconstruction of Small Moving Objects. *IEEE Transactions on Image Processing*, 19(11), 2901–2912.
63. Sajjadi, M., Scholkopf, B., Hirsch, M. (2017). Enhancenet: Single image super-resolution through automated texture synthesis. *In Proc. ICCV*, pp.4491–4500.
64. Noor, D. F., Li, Y., Li, Z., Bhattacharyya, S., York, G. (2019). Multi-Scale Gradient Image Super-Resolution for Preserving SIFT Key Points in Low-Resolution Images. *Signal Processing: Image Communication*, 78:236–245.
65. Huang, Y., Shao, L., Frangi, A. (2017). Simultaneous Super-Resolution and Cross-Modality Synthesis of 3D Medical Images using Weakly-Supervised Joint Convolutional Sparse Coding. *In Proc. CVPR*, pp.6070–6079.
66. You, C., Li, G., Zhang, Y., Zhang, X., Shan, H., Li, M., Ju, S., Zhao, Z., Zhang, Z., Cong, W., Vannier, M., Saha, P., Hoffman, E., Wang, G. (2020). CT Super-Resolution GAN Constrained by the Identical, Residual, and Cycle Learning Ensemble (GAN-CIRCLE). *IEEE Transactions on Medical Imaging*, 39(1), 188–203.
67. Tatem, A. J., Lewis, H. G., Atkinson, P. M., Nixon, M. S. (2001). Super-resolution target identification from remotely sensed images using a Hopfield neural network. *IEEE Transactions on Geoscience and Remote Sensing*, 39(4), 781–796.
68. Lei, S., Shi, Z., Zou, Z. (2020). Coupled Adversarial Training for Remote Sensing Image Super-Resolution. *IEEE Transactions on Geoscience and Remote Sensing*, 58(5), 3633–3643.

69. Choi, K., Kim, C., Kang, M., Ra, J. (2011). Resolution improvement of infrared images using visible image information. *IEEE Transactions on Image Processing*, 18(10), 611–614.
70. Han, T., Kim, D., Lee, S., Song, B. (2018). Resolution improvement of infrared images using visible image information. *Journal of Visual Communication and Image Representation*, 51:191–200.
71. Huang, J., Ma, L., Tan, T., Wang, Y. (2003). Learning based resolution enhancement of iris images. *In Proc. BMVC*, 1:1–10.
72. Yuan, Z., Wu, J., Kamata, S., Ahrary, A., Yan, P. (2009). Fingerprint image enhancement by super resolution with early stopping. *In Proc. ICIS*, 4:527–531.
73. Bian, W., Ding, S., Xue, Y. (2017). Fingerprint image super resolution using sparse representation with ridge pattern prior by classification coupled dictionaries. *IET Biometrics*, 6(5), 342–350.
74. Lin, F., Fookes, C., Chandran, V., Sridharan, S. (2005). Investigation into Optical Flow Super-Resolution for Surveillance Applications. *In Proc. APRS Workshop on Digital Image Computing*, pp.73–78.
75. Zhang, K., Zuo, W., Zhang, L. (2018a). Learning a single convolutional super-resolution network for multiple degradations. *In Proc. CVPR*, pp.3262–3271.
76. Sajjadi, M., Vemulapalli, R., Brown, M. (2018). Frame-Recurrent Video Super-Resolution. *In Proc. CVPR*, pp.6626–6634.
77. Daithankar, M. V., Ruikar, S. D. (2020). Video super resolution: A review. *In Proc. first Int. Conf. Data Science, Machine Learning and Applications*, 601:488–495.
78. Picku, L. C. (2007). Machine Learning in Multi-frame Image Super-resolution. *Ph.D thesis, University of Oxford, Britain*.
79. Mudenagudi, U., Banerjee, S., Kalra, P. K. (2011). Space-Time Super-Resolution Using Graph-Cut Optimization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(5), 995–1008.
80. Li, X., Hu, Y., Gao, X., Tao, D., Ning, B. (2010). A multi-frame image super-resolution method. *Signal Processing*, 90:405–414.
81. Li, K., Zhu, Y., Yang, J., Jiang, J. (2016). Video super-resolution using an adaptive superpixel-guided auto-regressive model. *Pattern Recognition*, 51:59–71, 2016.
82. Isobe, T., Li, S., Jia, X., Yuan, S., Slabaugh, G., Xu, C., Li, Y., Wang, S., Tian, Q. (2020). Video Super-Resolution With Temporal Group Attention. *In Proc. CVPR*, pp.8008–8017.
83. Liu, D., Wang, Z., Fan, Y., Liu, X., Wang, Z., Chang, S., Wang, X., Huang, T. (2018). Learning Temporal Dynamics for Video Super-Resolution: A Deep Learning Approach. *IEEE Transactions on Image Processing*, 27(7), 3432–3445.
84. Su, H., Wu, Y., Zhou, J. (2012). Super-Resolution Without Dense Flow. *IEEE Transactions on Image Processing*, 21(4), 1782–999.
85. Schoenemann, T., Cremers, D. (2012). A Coding-Cost Framework for Super-Resolution Motion Layer Decomposition. *IEEE Transactions on Image Processing*, 21(3), 1097–1110.
86. Babacan, S. D., Molina, R., Katsaggelos, A. K. (2011). Variational Bayesian Super Resolution. *IEEE Transactions on Image Processing*, 20(4), 984–999.
87. Dai, Q., Yoo, S., Kappeler, A., Katsaggelos, A. K. (2017). Sparse Representation-Based Multiple Frame Video Super-Resolution. *IEEE Transactions on Image Processing*, 26(2), 765–781.
88. Tu, Z., Xie, W., Zhang, D., Poppe, R., Veltkamp, R., Li, B., Yuan, J. (2019). A survey of variational and CNN-based optical flow techniques. *Signal Processing: Image Communication*, 72:9–24.
89. Tu, Z., Aa, Nico., Gemeren, C., Veltkamp, R. (2014). A combined post-filtering method to improve accuracy of variational optical flow estimation. *Pattern Recognition*, 47(5), 1926–1940.
90. Anwar, S., Khan, S., Barnes, N. (2020). A Deep Journey into Super-resolution: A Survey. *ACM Computing Surveys*, 53(3), 60:1–34.
91. Wang, L., Guo, Y., Liu, L., Lin, Z., Deng, X., An, W. (2020b). Deep Video Super-Resolution using HR Optical Flow Estimation. *IEEE Transactions on Image Processing*, 29:4323–4336.
92. Huang, Y., Wang, W., Wang, L. (2018). Video Super-Resolution via Bidirectional Recurrent Convolutional Networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(4), 1015–1028.
93. Liu, H., Ruan, Z., Zhao, P., Shang, F., Yang, L., Liu, Y. (2020). Video Super Resolution Based on Deep Learning: A comprehensive survey. *arXiv:2007.12928*.
94. Mitzel, D., Pock, T., Schoenemann, T., Cremers, D. (2009). Video super resolution using duality based TV-L1 optical flow. *In Proc. DAGM 2009: Pattern Recognition*, pp.432–441.
95. Baker, S., Kanade, T. (1999). Super-resolution optical flow. *Carnegie Mellon University, Pittsburgh, USA*, p.99–36.
96. Zhao, W., Sawhney, H. (2002). Is super-resolution with optical flow feasible? *In Proc. ECCV*, pp.599–613.

97. Keller, S., Lauze, F., Nielsen, M. (2011). Video Super-Resolution Using Simultaneous Motion and Intensity Calculations. *IEEE Transactions on Image Processing*, 20(7), 1870–1884.
98. Papenberg, N., Bruhn, A., Brox, T., Didas, S., Weickert, J. (2006). Highly accurate optic flow computation with theoretically justified warping. *International Journal of Computer Vision*, 67(2), 141–158.
99. Ma, Z., Liao, R., Tao, X., Xu, L., Jia, J., Wu, E. (2015). Handling motion blur in multi-frame super-resolution. *In Proc. CVPR*, pp.5224–5232.
100. Xiao, Z., Fu, X., Huang, J., Cheng, Z., Xiong, Z. (2021). Space-Time Distillation for Video Super-Resolution. *In Proc. CVPR*, pp.2113–2122.
101. Sun, D., Roth, S., Black, M. (2010). Secrets of optical flow estimation and their principles. *In Proc. CVPR*, pp.2432–2439.
102. Yang, W., Feng, J., Xie, G., Liu, J., Guo, Z., Yan, S. (2018). Video super-resolution based on spatial-temporal recurrent residual networks. *Computer Vision and Image Understanding*, 168:79–92.
103. Jo, Y., Oh, S., Kang, J., Kim, S. (2018). Deep video super-resolution network using dynamic upsampling filters without explicit motion compensation. *In Proc. CVPR*, pp.3224–3232.
104. Lucas, A., Lopez-Tapia, S., Molina, R., Katsaggelos, A. K. (2019). Generative Adversarial Networks and Perceptual Losses for Video Super-Resolution. *IEEE Transactions on Image Processing*, 28(7), 3312–3327.
105. Haris, M., Shakhnarovich, G., Ukita, N. (2019). Recurrent Back-Projection Network for Video Super-Resolution. *In Proc. CVPR*, pp.3897–3906.
106. Chan, K., Wang, X., Yu, K., Dong, C., Loy, C. (2021). BasicVSR: The Search for Essential Components in Video Super-Resolution and Beyond. *In Proc. CVPR*, pp.4947–4956.
107. Xiang, X., Tian, Y., Zhang, Y., Fu, Y., Allebach, J., Xu, C. (2020). Zooming Slow-Mo: Fast and accurate one-stage space-time video super-resolution. *In Proc. CVPR*, pp.3370–3379.
108. Xu, G., Xu, J., Li, Z., Wang, L., Sun, X., Cheng, M. (2021). Temporal Modulation Network for Controllable Space-Time Video Super-Resolution. *In Proc. CVPR*, pp.6388–6397.
109. He, K., Zhang, X., Ren, S., Sun, J. (2016). Deep Residual Learning for Image Recognition. *In Proc. CVPR*, pp.770–778.
110. Drulea, M., Nedeveschi, S. (2011). Total variation regularization of local global optical flow. *In Proc. ITSC*, pp.318–323.
111. Kim, T., Sajjadi, M., Hirsch, M., Scholkopf, B. (2018). Spatio-temporal transformer network for video restoration. *In Proc. ECCV*, pp.111–127.
112. Ronneberger, O., Fischer, P., Brox, T. (2015). U-net: Convolutional networks for biomedical image segmentation. *In Proc. MICCAI*, pp.234–241.
113. Xue, T., Chen, B., Wu, J., Wei, D., Freeman, W. (2019). Video enhancement with task-oriented flow. *International Journal of Computer Vision*, 127(8), 1106–1125.
114. Ranjan, A., Black, M. (2017). Optical flow estimation using a spatial pyramid network. *In Proc. CVPR*, pp.4161–4170.
115. Wang, Z., Yi, P., Jiang, K., Jiang, J., Han, Z., Lu, T., Ma, J. (2019). Multi-Memory Convolutional Neural Network for Video Super-Resolution. *IEEE Transactions on image processing*, 28(5), 2530–2544.
116. Bao, W., Lai, W., Zhang, X., Gao, Z., Yang, M. (2019). MEMC-Net: Motion estimation and motion compensation driven neural network for video interpolation and enhancement. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, DOI:10.1109/TPAMI.2019.2941941.
117. Lim, B., Son, S., Kim, H., Nah, S., Lee, K. M. (2017). Enhanced deep residual networks for single image super-resolution. *In Proc. CVPRW*, pp.1132–1140.
118. Dosovitskiy, A., Fischer, P., Ilg, E., Husser, P., Hazirbas, C., Golkov, V., Smagt, P., Cremers, D., Brox, T. (2015). FlowNet: Learning optical flow with convolutional networks. *In Proc. ICCV*, pp.2758–2766.
119. Kalarot, R., Porikli, F. (2019). MultiBoot VSR: Multistage multi-reference bootstrapping for video superresolution. *In Proc. CVPRW*, pp.2060–2069.
120. 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 Proc. CVPR*, pp.2462–2470.
121. Li, F., Bai, H., Zhao, Y. (2020). Learning a Deep Dual Attention Network for Video Super-Resolution. *IEEE Transactions on Image Processing*, 29:4474–4488.
122. Tao, X., Gao, H., Liao, R., Wang, J., Jia, J. (2017). Detail-revealing deep video super-resolution. *In Proc. ICCV*, pp.4472–4480.
123. Yi, P., Wang, Z., Jiang, K., Jiang, J., Lu, T., Tian, X., Ma, J. (2021). Omniscient Video Super-Resolution. *In Proc. ICCV*, pp.4429–4438.

124. Yang, X., Xiang, W., Zeng, H., Zhang, L. (2021). Real-world Video Super-resolution: A Benchmark Dataset and A Decomposition based Learning Scheme. *In Proc. ICCV*, pp.4781–4790.
125. Pan, J., Bai, H., Dong, J., Zhang, J., Tang, J. (2021). Deep Blind Video Super-resolution. *In Proc. ICCV*, pp.4811–4820.
126. Wang, W., Zhang, H., Yuan, Z., Wang, C. (2021b). Unsupervised Real-World Super-Resolution: A Domain Adaptation Perspective. *In Proc. ICCV*, pp.4318–4327.
127. Glorot, X., Bengio, Y. (2010). Understanding the difficulty of training deep feedforward neural networks. *In Proc. International Conference on Artificial Intelligence and Statistics*, pp.249–256.
128. Haris, M., Shakhnarovich, G., Ukita, N. (2018). Deep back-projection networks for super-resolution. *In Proc. CVPR*, pp.1664–1673.
129. Khrulkov, V., Babenko, A. (2021). Neural Side-By-Side: Predicting Human Preferences for No-Reference Super-Resolution Evaluation. *In Proc. CVPR*, pp.4988–4997.
130. Wang, Z., Bovik, A. C., Sheikh, H. R., Simoncelli, E. P. (2004). Image quality assessment: From error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4), 600–612.
131. Yang, W., Zhang, X., Tian, Y., Wang, W., Xue, J. H., Liao, Q. (2019). Deep learning for single image super-resolution: A brief review. *IEEE Transactions on Multimedia*, 21(12), 3106–3121.
132. Timofte, R., Gu, S., Wu, J., Van Gool, L., et al. (2018). Ntire 2018 challenge on single image super-resolution: Methods and results. *In Proc. CVPRW*, pp.965–976.
133. Ledig, C., Theis, L., Huszar, F., Caballero, J., Cunningham, A., Acosta, A., Aitken, A., Tejani, A., Totz, J., Wang, Z., Shi, W. (2017b). Photorealistic single image super-resolution using a generative adversarial network. *In Proc. CVPR*, pp.4681–4690.
134. Blau, Y., Mechrez, R., Timofte, R., Michaeli, T., Zelnik-Manor, L. (2018). The 2018 pirm challenge on perceptual image super-resolution. *In Proc. ECCVW*, pp.334–355.
135. Zhang, R., Isola, P., Efros, A. A., Shechtman, E., Wang, O. (2018b). The unreasonable effectiveness of deep features as a perceptual metric. *In Proc. CVPR*, pp.586–595.
136. Prashnani, E., Cai, H., Mostofi, Y., Sen, P. (2018). PieAPP: Perceptual image-error assessment through pairwise preference. *In Proc. CVPR*, pp.1808–1817.
137. Tang, L., Sun, K., Liu, L., Wang, G., Liu, Y. (2019). A reduced-reference quality assessment metric for super-resolution reconstructed images with information gain and texture similarity. *Signal Processing: Image Communication*, 79:32-39.