SAVSR: Arbitrary-Scale Video Super-Resolution via a Learned Scale-Adaptive Network

Zekun Li[1], Hongying Liu[2, 4*], Fanhua Shang[3*], Yuanyuan Liu[1*], Liang Wan[2, 3], Wei Feng[3]

[1]School of Artificial Intelligence, Xidian University, China
[2]Medical College, Tianjin University, Tianjin, China
[3]College of Intelligence and Computing, Tianjin University, Tianjin, China
[4]Peng Cheng Lab, Shenzhen, China
hyliu2009@tju.edu.cn; fhshang@tju.edu.cn; yyliu@xidian.edu.cn

Abstract

Deep learning-based video super-resolution (VSR) networks have gained significant performance improvements in recent years. However, existing VSR networks can only support a fixed integer scale super-resolution task, and when we want to perform VSR at multiple scales, we need to train several models. This implementation certainly increases the consumption of computational and storage resources, which limits the application scenarios of VSR techniques. In this paper, we propose a novel **Scale-adaptive Arbitrary-scale Video Super-Resolution network (SAVSR)**, which is the first work focusing on spatial VSR at arbitrary scales including both non-integer and asymmetric scales. We also present an omni-dimensional scale-attention convolution, which dynamically adapts according to the scale of the input to extract inter-frame features with stronger representational power. Moreover, the proposed spatio-temporal adaptive arbitrary-scale upsampling performs VSR tasks using both temporal features and scale information. And we design an iterative bi-directional architecture for implicit feature alignment. Experiments at various scales on the benchmark datasets show that the proposed SAVSR outperforms state-of-the-art (SOTA) methods at non-integer and asymmetric scales. The source code is available at `https://github.com/Weepingchestnut/SAVSR`.

1 Introduction

Super-resolution (SR) is the process of recovering high-resolution (HR) images or video frames from low-resolution (LR) ones, which is a fundamental problem in low-level vision tasks. Nowadays, streaming data represented by video is gradually becoming the mainstream of visual information carriers, but the mismatch between different video resolutions and numerous display devices has become a major concern. For example, 720P video does not provide satisfactory visual experience when played back on a 1080P or even 2K display device. Therefore, the study of video super-resolution (VSR) techniques has become one of the current research spotlights in low-level vision (Liu et al. 2022). In recent years, the rapid development of deep learning has greatly contributed to the progress of video super-resolution,

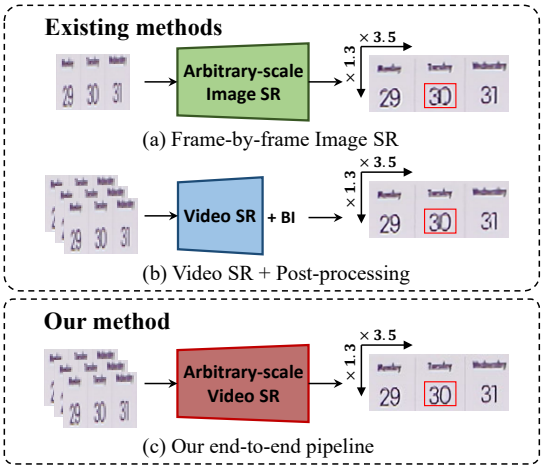

Figure 1: Comparison of the three ways for arbitrary-scale video super-resolution. (a) Frame-by-frame arbitrary-scale image super-resolution. (b) Video super-resolution and interpolation post-processing to resize the target resolution. (c) Our end-to-end arbitrary-scale video super-resolution. Note that "+BI" means that the network output is further resized to the target resolution by using the Bicubic interpolation.

and many deep learning-based methods (Liao et al. 2015; Caballero et al. 2017; Jo et al. 2018; Wang et al. 2019; Haris, Shakhnarovich, and Ukita 2019; Tian et al. 2020; Wang et al. 2020; Isobe et al. 2020b; Li et al. 2020; Isobe et al. 2020a; Chan et al. 2021; Liu et al. 2021; Yi et al. 2021; Yu et al. 2022; Wen et al. 2022; Lu et al. 2023) have been developed to improve the performance of VSR tasks.

Although deep learning-based VSR methods can obtain high performance, almost all the methods only consider certain specific integer scale factors (e.g., $\times 2$, $\times 3$, $\times 4$), or even develop models only for a scaling factor of 4. However, the scale requirements for VSR in practical applications may not be integers. For instance, when we want to super-resolve 480P video to 720P or 1080P, we need to enlarge $\times 1.5$ and $\times 2.25$, respectively. And the 1080P video should be super-resolved $\times 1.33$ to 2K videos. With the varying of current photography and display devices as well as the rise of video editing (Bar-Tal et al. 2022; Kasten et al. 2021) and art creation (Kasten et al. 2021; Siarohin et al. 2019; Ye and Bilodeau 2022), the arbitrary-scale VSR with asymmetric

*Corresponding authors

and non-integer scale factors has a wide range of applications. However, the upscale modules of existing VSR methods can only support a specific integer-scale SR task due to the usage of fixed filter-based sub-pixel convolution (Caballero et al. 2017), and they treat VSR with different scale factors as independent tasks, and therefore can not achieve arbitrary-scale VSR in real-world scenarios.

In single image super-resolution (SISR), there are several research works such as (Hu et al. 2019; Chen, Liu, and Wang 2021; Wang et al. 2021; Lee and Jin 2022; Son and Lee 2021; Lee, Choi, and Jin 2022) have utilized one model to realize arbitrary-scale SISR tasks. While for VSR, there are three ways for arbitrary-scale super-resolution. As shown in Fig. 1 (a), a simple solution is to take each frame in the video and adopt an arbitrary-scale SISR method to implement arbitrary-scale VSR frame by frame. However, this method ignores the temporal consistency between frames and easily introduces artefacts and jamming, leading to a worse visual experience. In contrast, the VSR methods can extract more feature information from the temporal dimension and generate HR videos with more natural detail and fewer artefacts. Therefore, in Fig. 1 (b), another naive idea is to firstly implement a specific scale of VSR (e.g. $\times 4$ factor) by using existing VSR methods, and then post-processing the resolution by traditional interpolation algorithms, i.e. a two-stage implementation of arbitrary-scale VSR. However, the drawbacks of such method can not be ignored. On the one hand, when we use interpolation post-processing, a certain amount of noise and blurring will be inevitably introduced. On the other hand, this two-stage pipeline increases the inference time, especially when the resolution of the video frames to be processed is high. Our work aims to realize an end-to-end arbitrary-scale VSR framework as shown in Fig. 1 (c), which achieves non-integer scale or even asymmetric scale video super-resolution.

In this paper, we propose a novel arbitrary-scale VSR network that achieves arbitrary-scale super-resolution requirements with only a single model. Specifically, we attempt to integrate the advantages of both iterative and recurrent architectures to propose an iterative bi-directional architecture (IBA). For arbitrary-scale requirements, small scales and higher resolution LR frame inputs are common application scenarios. Our IBA can improve the long-range modeling capability of the network with friendly computational resources. In addition, we introduce an omni-dimensional scale-attention convolution (OSConv) and a spatio-temporal adaptive arbitrary-scale upsampling (STAU) module. After the feature aggregation from multiple frames, the scale correlation in spatio-temporal features can be adaptively extracted by introducing OSConv in our network, which facilitates its adaptation to specific scale factors. The STAU dynamically generates upsampling convolution kernels based on the spatio-temporal information of the input video clips, by adding the scale information, and maps the LR features into the HR space to realize arbitrary-scale upsampling.

To the best of our knowledge, our model is the first work to concentrate on spatial VSR at arbitrary scales, including non-integer, asymmetric scales. Our main contributions can be summarized as follows:

- We propose a novel iterative bi-directional VSR architecture, which can effectively improve the long-range modeling capability of the proposed network with the same memory consumption as an iterative architecture.
- We present omni-dimensional scale-attention convolution and spatio-temporal adaptive arbitrary-scale upsampling, which modify the weights according to the input scale information, in order to achieve non-integer, asymmetric scale VSR with better adaptive capability.
- Various experimental results show that our method can achieve excellent arbitrary-scale VSR performance on both non-integer and asymmetric scales, and also competitive performance with other methods for common integer scales using only one model.

2 Related Work

In this section, we briefly review the progress of arbitrary-scale SISR, and introduce several works on VSR. And we discuss adaptive convolutions related to this work.

Arbitrary-scale Image Super-Resolution. Some SISR methods such as (Dong et al. 2014; Kim, Lee, and Lee 2016; Shi et al. 2016; Mao, Shen, and Yang 2016; Lim et al. 2017; Ledig et al. 2017; Zhang et al. 2018b,a; Liang et al. 2021) based on deep convolutional neural networks can achieve remarkable performance. However, conventional methods only consider the SR of certain integer scale factors, ignoring the non-integer scale needs for arbitrary-scale SR in real-world scenes (Liu et al. 2023). Some earlier methods that use traditional interpolation to upsample, such as VDSR (Kim, Lee, and Lee 2016), can easily achieve multi-scale SR, but the limitations of interpolation easily lead to poor performance. Meta-SR (Hu et al. 2019) was proposed to implement arbitrary-scale SISR with a single model for the first time by using meta-learning to dynamically predict, rather than store the filter weights at each scale. ArbSR (Wang et al. 2021) designs two plug-ins, the scale-aware feature adaptation block and the scale-aware upsampling layer, that further enable existing SISR networks to implement asymmetric, non-integer-scale SR. LIIF (Chen, Liu, and Wang 2021) introduces implicit neural representation (Mildenhall et al. 2021) to SISR, thus realizing arbitrary resolution image representation. Subsequent works are also based on the idea of implicit representation to further improve the performance of arbitrary-scale SISR (Lee and Jin 2022; Lee, Choi, and Jin 2022; Ma et al. 2022) or to extend it to new application areas (Yang et al. 2021). However, few works have focused on arbitrary-scale VSR.

Video Super-Resolution. The architectures of existing VSR networks can be broadly classified into iterative and recurrent methods. The networks based on iterative architectures usually take multiple LR frames as inputs and output SR result for the middle frame. For instance, DUF-VSR (Jo et al. 2018) designs a 3D convolutional network with dynamic upsampling filters. EDVR (Wang et al. 2019) proposes a multi-scale pyramid structure, a deformable alignment module and a spatio-temporal attention fusion module. STAN (Wen et al. 2022) introduces an adaptive filter

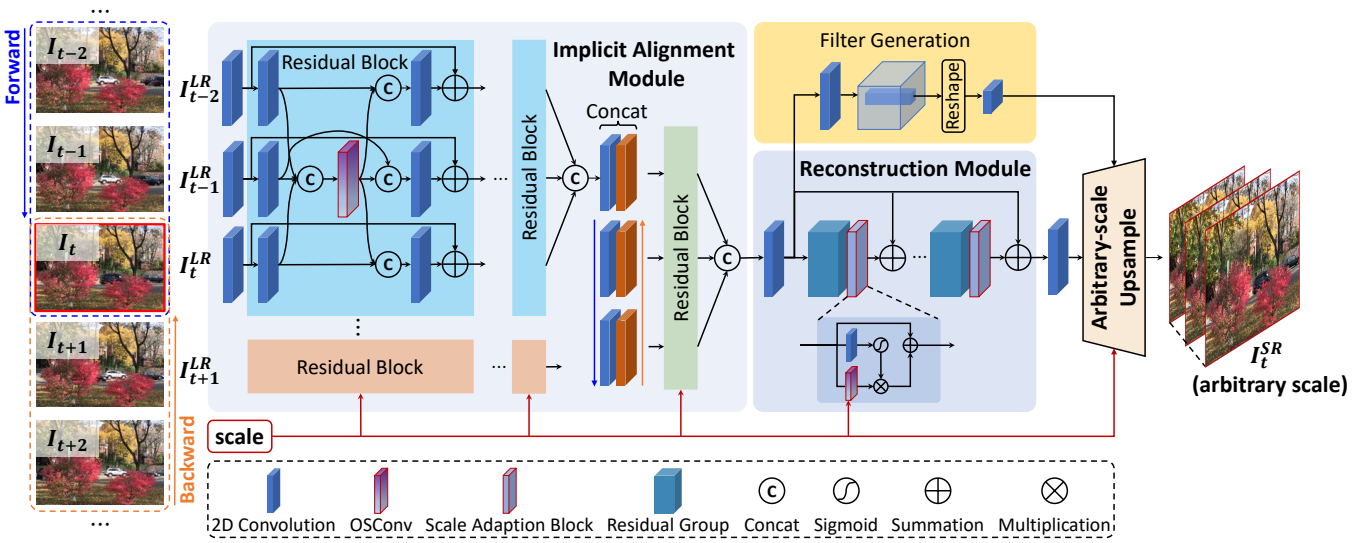

Figure 2: The overall architecture of the proposed SAVSR method. Note that here scale is a tuple containing horizontal and vertical scale factors r_h, r_v. OSConv denotes the proposed omni-dimensional scale-adaptive convolution. The residual blocks with blue and orange backgrounds correspond to forward and backward feature extraction, respectively, and the green residual block represents the aggregation of both forward and backward features.

for spatio-temporal alignment. Iterative architecture-based models are relatively small in size because they share parameters across frames, but they tend to have insufficient temporal feature extraction. Recurrent architectures are generally multi-frame input and multi-frame output, processing all frames in parallel. For example, BasicVSR (Chan et al. 2021) uses a bidirectional propagation scheme to maximise the extraction of temporal information. However, these models are always large in size and require a large amount of GPU memory for inference. In light of both works, our SAVSR attempts to combine the advantages of both iterative and recurrent frameworks while maintaining a balance between model size and performance.

Dynamic Convolutions. Conventional convolution only has static filters, and the filter weights are independent of input samples. The dynamic convolution performs linear weighting of multiple filters with input-dependent weights, and thus has a dynamic nature that varies according to the input samples. Some works (Yang et al. 2019; Chen et al. 2020; Ma et al. 2020; Li, Zhou, and Yao 2022) show that dynamic convolution has stronger representation capability than common convolution, and can effectively increase the network capacity without adding extra computational cost. Therefore, we use scale factors and spatio-temporal features as the basis for generating dynamic weights, and propose spatio-temporal scale-adaptive convolution to achieve better arbitrary-scale VSR performance than counterparts.

3 Methodology

In this section, we present in detail our arbitrary-scale video super-resolution network SAVSR shown in Fig. 2.

3.1 Network Overview

In fact, our SAVSR is an iterative architecture, which takes $2N+1$ consecutive LR frames and the currently requested

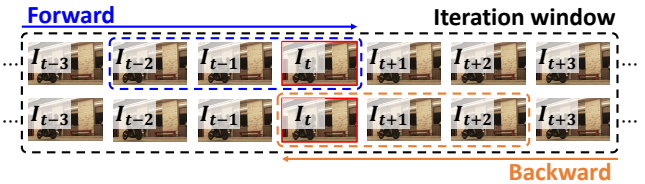

Figure 3: Our iterative bi-directional architecture.

horizontal/vertical scale factors r_h, r_v as inputs, and outputs super-resolved intermediate frames I_t^{SR} at the corresponding scale. The overall objective function of our method is

$$I_t^{SR} = \mathcal{F}_{SAVSR}(I^{LR}_{[t-N:t+N]}, s), \tag{1}$$

where $\mathcal{F}_{SAVSR}$ denotes our SAVSR model, $I^{LR}_{[t-N:t+N]}$ is the $2N+1$ consecutive LR frames, and s is a tuple, i.e., $s = (r_h, r_v), r_h, r_v \in \mathbb{R}$.

Our SAVSR model consists of three main components: an implicit alignment module, a reconstruction module, and arbitrary-scale upsampling. The implicit alignment module performs temporal feature extraction and alignment of the input continuous frames using an iterative bi-directional architecture, where each residual block contains an OSConv to adjust dynamically to the scale factors. The implicitly aligned feature maps are fed into the reconstruction module for feature refinement. The LR frames are simultaneously used for dynamic weight prediction for arbitrary-scale upsampling, aiming to introduce temporal information into the upsampling module. Finally, the arbitrary-scale upsampling converts the refined feature maps into SR frames at the current scale. Below, we describe each module of our SAVSR in detail.

3.2 Iterative Bi-directional Alignment Architecture

For arbitrary-scale VSR tasks, the input LR frames are generally of higher resolution when our target scale is small,

e.g., $\times 1.33$, $\times 1.5$. Although the recurrent architecture has the good capability to merge temporal information, it requires huge computational resources and storage capacity. While the iterative architecture is resource-friendly but lacks excellent temporal modeling capabilities. Therefore, we propose an iterative bi-directional architecture (IBA) for our arbitrary-scale VSR task that attempts to combine the advantages of the above architectures. It is a bi-directional recurrent architecture within an iterative window shown in Fig. 3. From a global perspective, IBA is still an iterative architecture that takes multiple consecutive LR frames $I^{LR}_{[t-N:t+N]}$ as inputs and outputs the super-resolved center reference frame I^{SR}_t. Different from general iterative architectures, we utilize a small sliding window of size 3 in each iteration to extract inter-frame information in both directions for implicit alignment. Therefore, we exploit all the frame information during an iteration, which is similar to recurrent architectures. We set the direction from the past to the future as forward propagation and the opposite as backward. Such a bi-directional design enables the model to take full advantage of temporal information within an iteration window, establishing the basis for subsequent implicit alignment.

3.3 Omni-dimensional Scale-attention Convolution

Super-resolution tasks at different scales are not independent, but correlated. Some studies (Hu et al. 2019; Wang et al. 2021; Zhang, Gool, and Timofte 2020) indicate that super-resolution models have certain similarities and differences at both the filter and feature levels when faced with different scale super-resolution tasks. Therefore, how to exploit these similarities and differences according to specific scales is the key to improving a single model for arbitrary-scale super-resolution tasks.

To better represent feature relevance based on scale information, we propose an omni-dimensional scale-attention convolution (OSConv) that adaptively modifies the weights of the convolution based on the current scale factors. Fig. 4 shows the structure of our OSConv. Given an input feature map $X \in \mathbb{R}^{C\times H\times W}$ and the scale factors r_h, r_v. We want to tailor the adaptive weights for the convolutional kernels using the current scale in four dimensions, namely the spatial kernel size $k \times k$, the number of input channels c_{in} and the number of output channels c_{out} for each convolutional kernel, and the number of convolutional kernels n. First, we concatenate and feed the inverse of the horizontal and vertical scale factors into a Multi-layer Perceptron (MLP) that maps the scale information into a feature vector of length C, expressed by the following equation:

$$F_{scale} = \mathrm{MLP}(\mathrm{concat}(\frac{1}{r_h}, \frac{1}{r_v}, GAP(X))), \qquad (2)$$

where $F_{scale} \in \mathbb{R}^{C\times 1\times 1}$ is the scale vector. Then, we use the scale feature vector to learn the scale attention of the convolution kernels from 4 dimensions, which is embodied by the four attention branches containing the fully connected (FC) layers. Specifically, for each attention branch, the feature vector is passed through the FC layer to produce attention weights of size $k\times k$, $c_{in}\times 1$, $c_{out}\times 1$, and $n\times 1$, respectively, and normalized with a Sigmoid or Softmax function.

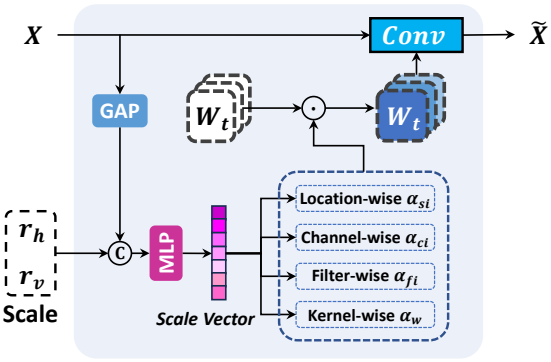

Figure 4: The details of our omni-dimensional scale-attention convolution (OSConv). "MLP" denotes multi-layer perceptron and "GAP" is global average pooling. "$\odot$" denotes the multiplication operations along different dimensions of the kernel space.

Finally, we assign the attention scalar to the convolutional kernel weights and output the feature map $\tilde{X}$ as follows:

$$\begin{aligned}\tilde{X} = (&\alpha_{w1} \odot \alpha_{f1} \odot \alpha_{c1} \odot \alpha_{s1} \odot W_1 + ... \\ &+ \alpha_{wn} \odot \alpha_{fn} \odot \alpha_{cn} \odot \alpha_{sn} \odot W_n) * X,\end{aligned} \qquad (3)$$

where $\alpha_{wi} \in \mathbb{R}$, $\alpha_{si} \in \mathbb{R}^{k\times k}$, $\alpha_{ci} \in \mathbb{R}^{c_{in}}$ and $\alpha_{fi} \in \mathbb{R}^{c_{out}}$ denotes the attention scalar of the convolution kernel W_i, the spatial dimension along the convolution kernel space, the input channel dimension and the output channel dimension, respectively. $\odot$ denotes the multiplication along different dimensions of the kernel space, and $*$ is convolution.

3.4 Spatio-Temporal Adaptive Arbitrary-scale Upsampling

The proposed spatio-temporal adaptive arbitrary-scale upsampling (STAU) is shown in Fig. 5. STAU has two branches, the spatio-temporal filter branch and the scale filter branch, the former generates spatio-temporal filters $\mathcal{F}_{stf}$ for adaptive convolution based on implicitly aligned features, and the latter generates scale filters $\mathcal{F}_{sf}$ for adaptive convolution based on current scale information. Finally, we aggregate the features of both branches as the output of the upsampling. Specifically, we assume that the coordinates of a pixel in the HR space are (x, y). We first calculate the coordinates $C(x), C(y)$ and the relative distance $R(x), R(y)$ of the pixel (x, y) projected into the LR space. For the horizontal coordinate x, we have

$$C(x) = \frac{x+0.5}{r_h} - 0.5, \;\; R(x) = C(x) - \mathrm{floor}(\frac{x+0.5}{r_h}). \qquad (4)$$

Similarly, $C(y)$ and $R(y)$ can be computed. Note that we follow the geometric setting of treating pixels as squares rather than points, so we calculate the geometric center of the pixel. Then we concatenate the inverse of the scale factors $\frac{1}{r_h}, \frac{1}{r_v}$ with the relative distance $R(x), R(y)$ and feed them into an MLP for feature transformation, and the transformed features generate the scale filters and two sets of offsets, respectively. The offsets are used for feature sampling, i.e., the domain features centered at coordinates

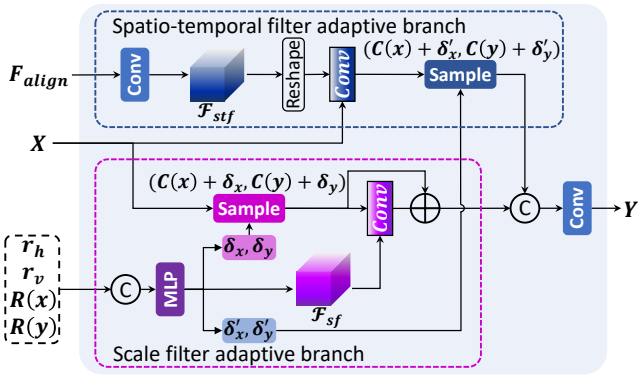

Figure 5: An illustration of our spatio-temporal adaptive arbitrary-scale upsampling (STAU).

$(C(x)+\delta_x, C(y)+\delta_y)$ are sampled using Bilinear interpolation. The sampled features tend to destroy the local spatio-temporal correlations in the LR domain. Therefore, for the spatio-temporal filter branch, we perform spatio-temporal filter adaptive convolution before sampling. For the scale filter branch, we first sample the LR features into the HR domain with offsets, and then refine the features with scale-adaptive filters. The spatio-temporal and scale adaptivity makes our upsampling more suitable for VSR.

4 Experiments

4.1 Datasets and Metrics

We use the training set from Vimeo-90K (Xue et al. 2019) dataset which contains over 9000 training and testing video sequences. Each video sequence includes 7 consecutive frames with resolution of 448×256. Bicubic (Keys 1981) is adopted for arbitrary-scale downsampling to generate LR frames with different resolutions. Moreover, we employ the benchmark datasets Vid4 (Liu and Sun 2011) and UDM10 (Yi et al. 2019) to evaluate our model. Peak-Signal-to-Noise Ratio (PSNR) and Structual Similarity Index (SSIM) are used as evaluation metrics. Following the arbitrary-scale SISR works (Hu et al. 2019; Wang et al. 2021; Chen, Liu, and Wang 2021), for some non-integer scales we crop the frame borders to ensure that downsampling these scales does not result in fractional resolution, e.g., a 576×704 frame must be cropped to 575×700 when downsampling by $\times 2.5$, to make the LR frame with an integer resolution 230×280. We calculate all the metrics in the luminance channel as that in the conventional VSR works.

4.2 Comparison of Methods

Moreover, we compare our SAVSR with two classes of existing methods for arbitrary-scale VSR, as shown in Fig. 1. (1) Frame-by-frame Image SR. We selected the representative methods: Meta-SR (Hu et al. 2019), LIIF (Chen, Liu, and Wang 2021), ArbSR (Wang et al. 2021) and LTE (Lee and Jin 2022), where LIIF, ArbSR and LTE can achieve asymmetric-scale super-resolution. For Meta-SR, we also add interpolation post-processing to achieve super-resolution at asymmetric scales. (2) Video SR + Post-processing. We choose DUF (Jo et al. 2018), EDVR (Wang et al. 2019), OVSR (Yi et al. 2021), BasicVSR (Chan et al. 2021) and VideoINR (Chen et al. 2022) for comparison, where VideoINR is a spatio-temporal super-resolution method that enables super-resolution on non-integer scales. Note that for VSR methods that can only achieve a single fixed integer scale, we select the smallest integer of super-resolution results larger than the current required scale for post-interpolation processing. For example, $\times 1.8$ and $\times 2.3$ SR are obtained from the results of $\times 2$ and $\times 3$ by using Bicubic downsampling, respectively.

The quantitative results for both asymmetric and symmetric-scale VSR are shown in Table 1. It is clear that the VSR methods outperform the arbitrary-scale image SR methods, which shows the importance of temporal information in VSR tasks. Compared with the state-of-the-art (SOTA) VSR methods that focus on a single integer scale, our SAVSR also achieves comparable performance. SAVSR obtains the highest PSNR at all randomly selected scale factors. For example, on UDM10 at $\frac{\times 1.5}{\times 3.5}$, PSNR is improved by 2.28 dB compared to the best-performing arbitrary-scale image SR method, ArbSR, and by 0.48 dB, 0.08 dB, and 0.36 dB than the VSR methods, EDVR, OVSR, and BasicVSR, respectively. The high performance of our SAVSR is because our network architecture is designed for arbitrary-scale VSR tasks, and both the implicit alignment module and the spatio-temporal adaptive upsampling make the advantage of video temporal information. Furthermore, the two adaptive filter branches of STAU allow the model to utilize both spatio-temporal and scale information in upsampling, which are not available in other arbitrary-scale upsamplings.

Fig. 6 compares the visual results of asymmetric and symmetric-scale super-resolution for the City video sequence on the Vid4 test set. It can be seen from zoom-in regions that the results of our SAVSR show better perceptual quality and fewer artifacts. As the SISR methods are capable of asymmetric scale tasks, the results of LIIF are too smooth without clear edges, while ArbSR produces the results that have sharp edges but introduce error messages that are inconsistent with the ground truth (GT). For VSR methods, since they have to adjust the resolution by interpolation post-processing, their results exhibit various artifacts. For example, they show blurred artifacts in dense rectangular panes. VideoINR tends to introduce more artifacts in the high-frequency region when fulfils the spatial VSR tasks. In contrast, our SAVSR shows clearer and finer details. Fig. 7 illustrates our method for arbitrary-scale VSR. More results are reported in the Supplementary Material.

4.3 Ablation Studies

The following ablation experiments on the Vid4 dataset investigate the importance of each of our proposed module. The training set for all ablation experiments is Vimeo90K.

The STAU and OSConv modules. The ablation studies of our SAVSR with STAU and OSConv modules are listed in Table 2. Note that in the experiment, each time we only introduce one module when comparing with the baseline model. Firstly, we demonstrate the effectiveness of temporal information for upsampling based on whether spatio-temporal filter branch is introduced. For symmetric

Method	PSNR (dB) / SSIM at Symmetric and Asymmetric Scale							
	Vid4				UDM10			
	×1.5	×2.2	×2.7	×3.5	×1.3	×2.3	×2.9	×3.7
Bicubic	32.23 / 0.9441	27.65 / 0.8442	25.94 / 0.7723	24.41 / 0.6794	44.84 / 0.9927	37.08 / 0.9602	34.80 / 0.9369	32.87 / 0.9085
Meta-SR (Hu et al. 2019) + BI	29.26 / 0.8835	25.92 / 0.7945	24.65 / 0.7077	23.46 / 0.6339	33.11 / 0.8016	30.64 / 0.8140	30.40 / 0.8314	28.19 / 0.7283
LIIF (Chen, Liu, and Wang 2021)	36.70 / 0.9755	30.76 / 0.9169	28.23 / 0.8618	26.30 / 0.7815	49.24 / 0.9971	42.16 / 0.9825	39.45 / 0.9687	37.04 / 0.9500
ArbSR (Wang et al. 2021)	36.41 / 0.9765	29.85 / 0.9160	27.29 / 0.8633	25.19 / 0.7821	43.85 / 0.9921	37.53 / 0.9781	35.48 / 0.9643	33.37 / 0.9446
LTE (Lee and Jin 2022)	36.77 / 0.9757	30.71 / 0.9170	28.32 / 0.8630	26.26 / 0.7823	49.07 / 0.9969	42.12 / 0.9826	39.45 / 0.9690	37.07 / 0.9503
DUF (Jo et al. 2018) + BI	36.91 / 0.9770	31.56 / 0.9382	29.76 / 0.9067	27.75 / 0.8449	OOM	43.60 / 0.9862	41.23 / 0.9770	38.79 / 0.9620
EDVR (Wang et al. 2019) + BI	37.71 / 0.9807	32.48 / 0.9444	29.65 / 0.9053	28.10 / 0.8557	49.68 / 0.9977	42.90 / 0.9883	41.53 / 0.9773	39.17 / **0.9693**
OVSR (Yi et al. 2021) + BI	37.17 / 0.9784	32.25 / 0.9480	30.02 / 0.9115	28.07 / 0.8582	50.38 / 0.9974	43.95 / 0.9874	41.57 / 0.9791	39.03 / 0.9650
BasicVSR (Chan et al. 2021) + BI	37.80 / 0.9810	32.45 / 0.9494	30.12 / 0.9189	27.81 / 0.8505	51.22 / 0.9979	44.34 / 0.9882	41.43 / 0.9804	39.35 / 0.9661
VideoINR (Chen et al. 2022)	26.52 / 0.8385	27.78 / 0.8755	27.60 / 0.8619	26.62 / 0.8093	30.58 / 0.9025	34.29 / 0.9494	32.51 / 0.9289	36.23 / 0.9528
SAVSR (Ours)	**38.00 / 0.9815**	**33.01 / 0.9497**	**30.40 / 0.9194**	**28.22 / 0.8599**	**52.14 / 0.9982**	**44.55 / 0.9891**	**42.20 / 0.9809**	**39.65** / 0.9671
	×2/×4	×1.6/×3.05	×3.9/×2	×3.5/×1.5	×1.5/×4	×1.6/×3.05	×3.5/×1.75	×4/×1.4
Bicubic	25.27 / 0.7360	27.05 / 0.8198	25.42 / 0.7493	26.39 / 0.8061	34.63 / 0.9615	36.41 / 0.9603	35.60 / 0.9413	35.00 / 0.9341
Meta-SR (Hu et al. 2019) + BI	24.28 / 0.6984	25.67 / 0.7666	24.06 / 0.6672	24.97 / 0.7367	30.82 / 0.8857	OOM	30.61 / 0.8667	30.66 / 0.8768
LIIF (Chen, Liu, and Wang 2021)	27.69 / 0.8318	29.72 / 0.8915	27.38 / 0.8350	28.49 / 0.8745	38.55 / 0.9705	41.00 / 0.9815	39.79 / 0.9691	38.73 / 0.9623
ArbSR (Wang et al. 2021)	27.57 / 0.8350	29.70 / 0.8942	25.85 / 0.8344	27.16 / 0.8766	35.39 / 0.9673	37.22 / 0.9774	35.43 / 0.9648	34.61 / 0.9582
LTE (Lee and Jin 2022)	27.75 / 0.8332	29.75 / 0.8925	27.39 / 0.8352	28.52 / 0.8755	38.53 / 0.9706	40.95 / 0.9815	39.80 / 0.9692	38.78 / 0.9625
DUF (Jo et al. 2018) + BI	29.19 / 0.8765	31.82 / 0.9296	28.73 / 0.8773	29.75 / 0.9034	40.60 / 0.9788	43.07 / 0.9866	41.28 / 0.9750	40.23 / 0.9687
EDVR (Wang et al. 2019) + BI	28.87 / 0.8693	31.58 / 0.9257	28.44 / 0.8721	29.42 / 0.9001	40.47 / 0.9798	43.37 / 0.9878	41.46 / 0.9781	40.39 / 0.9691
OVSR (Yi et al. 2021) + BI	29.64 / 0.8864	32.21 / 0.9334	28.96 / 0.8841	30.03 / 0.9111	40.84 / 0.9802	43.32 / 0.9876	41.52 / 0.9768	40.40 / 0.9705
BasicVSR (Chan et al. 2021) + BI	29.27 / 0.8791	32.15 / 0.9331	28.98 / 0.8847	29.97 / 0.9102	40.31 / 0.9805	43.36 / 0.9888	41.47 / 0.9789	40.21 / 0.9724
SAVSR (Ours)	**29.82 / 0.8881**	**32.43 / 0.9351**	**29.14 / 0.8873**	**30.23 / 0.9131**	**41.32 / 0.9812**	**43.88 / 0.9886**	**42.17 / 0.9795**	**40.77 / 0.9732**

Table 1: The super-resolution results for symmetric and asymmetric scale factors on datasets Vid4 and UDM10. "+BI" means downsampling SR to the target resolution using Bicubic interpolation post-processing.

Feature Backbone		STAU	Symmetric Scale								Asymmetric Scale			
OSConv in Alignment	OSConv in Reconstruction	Spatio-Temporal Filter Branch	×1.3	×1.8	×2	×2.25	×2.7	×3.3	×3.5	×4	$\frac{\times 1.9}{\times 3.5}$	$\frac{\times 1.3}{\times 3.9}$	$\frac{\times 3.35}{\times 1.6}$	$\frac{\times 3.3}{\times 2}$
✗	✗	✗	38.81	34.68	33.71	31.78	29.95	28.08	27.84	26.91	30.33	29.79	29.94	29.88
✗	✗	✓	39.22	34.89	33.83	31.90	30.10	28.19	28.03	27.01	30.58	29.93	30.22	29.95
✓	✗	✓	39.49	34.98	34.15	32.18	30.25	28.33	28.08	27.04	30.68	30.17	30.31	30.07
✗	✓	✓	39.55	35.14	34.24	32.24	30.28	28.37	28.11	27.07	30.73	30.21	30.30	30.10
✓	✓	✓	**39.77**	**35.21**	**34.44**	**32.30**	**30.40**	**28.48**	**28.22**	**27.19**	**30.83**	**30.29**	**30.37**	**30.17**

Table 2: The ablation studies of our SAVSR with OSConv and STAU modules.

OSConv		Scale	
Scale Vector	GAP	×3.5/×1.5	×3.5
✗	✗	30.01 (+0.00)	27.97 (+0.00)
✓	✗	30.18 (+0.17)	28.15 (+0.18)
✓	✓	30.23 (+0.22)	28.22 (+0.25)

Table 3: The ablations on attention scalars of our OSConv.

scale $\times 3.5$ and asymmetric scale $\frac{\times 1.9}{\times 3.5}$, the upsampling with the spatio-temporal filter adaptive convolution improves the PSNR results by > 0.19 and > 0.25 dB, respectively. For OSConv, we introduce it in different modules of the feature backbone to show its effectiveness. It can be seen that when OSConv is introduced into both alignment and reconstruction modules, the PSNR enhances higher than in the single case, which confirms our design. For instance, the PSNR is improved by 0.4 and 0.36 dB for symmetric scale $\times 2.25$ and asymmetric scale $\frac{\times 1.3}{\times 3.9}$, respectively, when OSConv is introduced in both modules. In summary, scale adaptation in the whole feature backbone can further enhance the representation of the network and improve the performance of the model. For the structure of the proposed OSConv, Table 3 indicates the effectiveness of scale attention.

As shown in Table 4, to further demonstrate the effectiveness of the proposed STAU for arbitrary-scale VSR tasks, we combine our feature backbone with the most representative arbitrary-scale upsampling currently. FLOPs are computed on one LR feature map with the size of $64 \times 180 \times 320$ and $\times 4$ upsampling. Our STAU module achieves optimal performance with fewer parameters and FLOPs.

In Fig. 8, we show the comparison of the temporal profiles between SAVSR and the SOTA method, BasicVSR. We demonstrate the horizontal and vertical temporal profiles of a region with the $\times 3.5$ VSR task. It can be seen that the profiles from BasicVSR have artifacts, which may be introduced due to BI post-processing. By our adaptive temporal alignment and STAU, the profiles from our SAVSR show a smoother transition.

Iterative Bi-directional Alignment Architecture Table 5 shows the ablations of the our iterative bi-directional alignment architecture. With the same iteration window and sliding window settings, bi-directional propagation is more advantageous, improving 0.43 and 0.44 dB over unidirectional propagation for the given symmetric and asymmetric scales, respectively. We further explore the optimal combination of

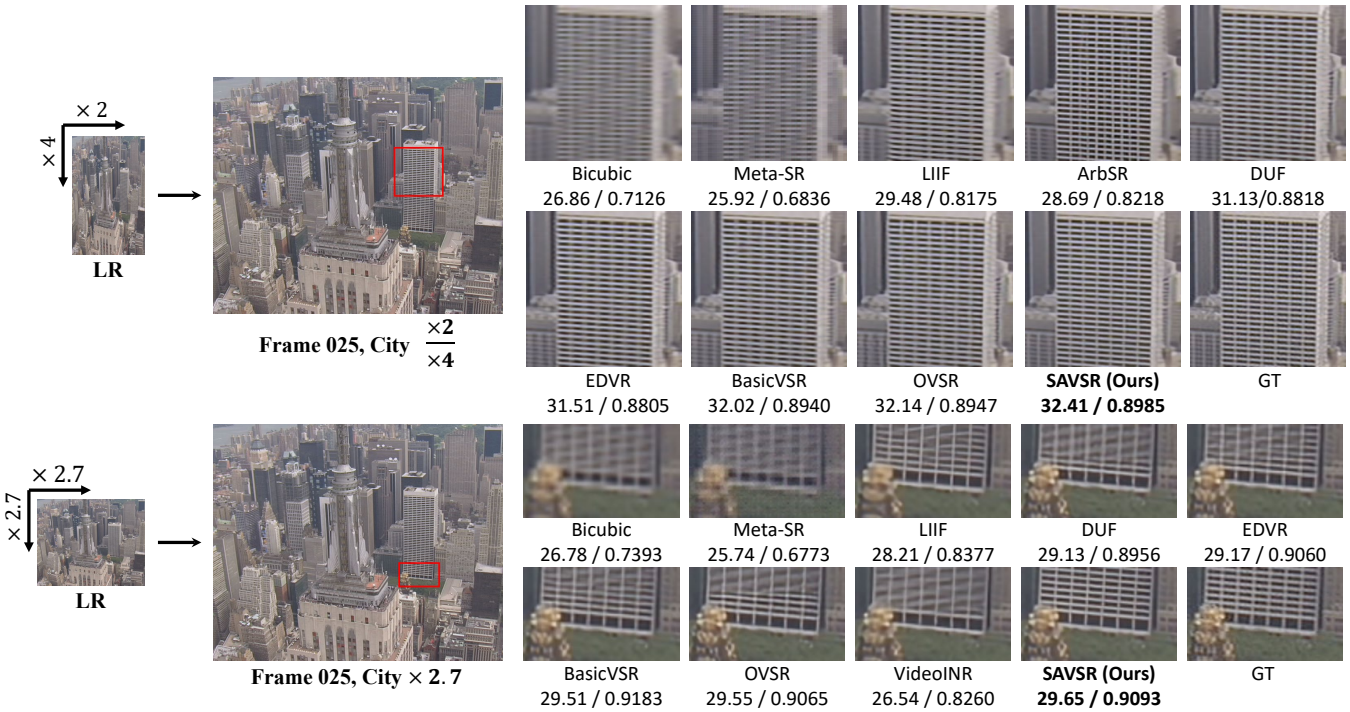

Figure 6: Visual comparison of the VSR methods (PSNR/SSIM) at symmetric and asymmetric scale factors on the Vid4 dataset.

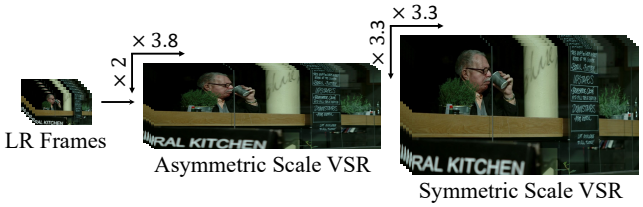

Figure 7: Examples of our SAVSR for arbitrary-scale VSR.

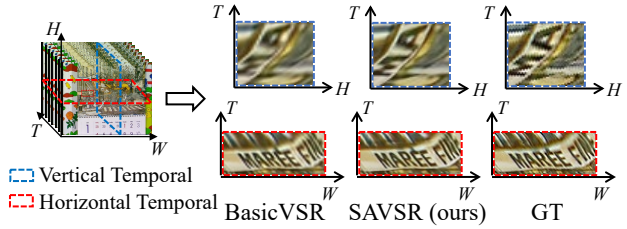

Figure 8: Comparison of temporal profile.

iterative window size and sliding window size. It can be seen from Table 5 that the expansion of the iteration and sliding windows is not necessarily proportional to the gain of PSNR. On the contrary, it increases the parameters of the model. Therefore, considering the trade-off between performance and complexity, we set the iteration window size to 7 and the sliding window size to 3 for the Vimeo90K dataset.

5 Conclusions

In this paper, we proposed a novel scale-adaptive video super-resolution network (SAVSR), which can deal with arbitrary-scale VSR tasks with only a single model. We enable the proposed network to adaptively modulate features according to the current scale by introducing omni-dimensional scale-attention convolution to super-resolution tasks at different scale factors. The proposed spatio-temporal adaptive arbitrary-scale upsampling can directly introduce the temporal information of alignment features to accomplish upsampling. We also designed an iterative bi-directional alignment architecture for implicit alignment in arbitrary-scale VSR. Extensive experimental results showed that our method can obtain promising results on VSR tasks at both non-integer and asymmetric scale factors, while it enjoys competitive performance for common integer scale factors. In the future, we will focus on more efficient techniques for general arbitrary-scale VSR tasks.

Upsampling	Params. (K)	FLOPs (G)	Scale	
			×3.7	×2.95/×3.75
Meta-Upscale	+ 445	+ 410	25.84	25.59
LIIF-Upscale	+ 347	+ 1283	26.73	26.71
LTE-Upscale	+ 494	+ 753	26.77	26.83
SATU (Ours)	**+ 121**	**+ 23**	**27.65**	**28.26**

Table 4: The comparison of our STAU module with other arbitrary-scale upsamplings.

Propagation	Params. (M)	Iteration WS	Slide WS	Scale	
				×2.25	×2/×3.2
Unidirectional	8.0	7	3	31.84	30.75
Bi-directional	11.5	7	3	32.27	31.19
	12.3	7	5	32.24	31.18
	16.8	13	3	32.30	31.20
	19.9	13	5	32.33	31.25

Table 5: The PSNR results of our SAVSR with the proposed iterative bi-directional alignment architecture on the Vid4 dataset. 'WS' means window size.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (Nos. 62276182, 61976164, and 62072334), and Natural Science Basic Research Program of Shaanxi (Program No. 2022GY-061).

References

Bar-Tal, O.; Ofri-Amar, D.; Fridman, R.; Kasten, Y.; and Dekel, T. 2022. Text2live: Text-driven layered image and video editing. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XV*, 707–723. Springer.

Caballero, J.; Ledig, C.; Aitken, A.; Acosta, A.; Totz, J.; Wang, Z.; and Shi, W. 2017. Real-time video super-resolution with spatio-temporal networks and motion compensation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 4778–4787.

Chan, K. C.; Wang, X.; Yu, K.; Dong, C.; and Loy, C. C. 2021. Basicvsr: The search for essential components in video super-resolution and beyond. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4947–4956.

Chen, Y.; Dai, X.; Liu, M.; Chen, D.; Yuan, L.; and Liu, Z. 2020. Dynamic convolution: Attention over convolution kernels. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 11030–11039.

Chen, Y.; Liu, S.; and Wang, X. 2021. Learning continuous image representation with local implicit image function. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8628–8638.

Chen, Z.; Chen, Y.; Liu, J.; Xu, X.; Goel, V.; Wang, Z.; Shi, H.; and Wang, X. 2022. Videoinr: Learning video implicit neural representation for continuous space-time super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2047–2057.

Dong, C.; Loy, C. C.; He, K.; and Tang, X. 2014. Learning a deep convolutional network for image super-resolution. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part IV 13*, 184–199. Springer.

Haris, M.; Shakhnarovich, G.; and Ukita, N. 2019. Recurrent back-projection network for video super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3897–3906.

Hu, X.; Mu, H.; Zhang, X.; Wang, Z.; Tan, T.; and Sun, J. 2019. Meta-SR: A magnification-arbitrary network for super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1575–1584.

Isobe, T.; Jia, X.; Gu, S.; Li, S.; Wang, S.; and Tian, Q. 2020a. Video super-resolution with recurrent structure-detail network. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XII 16*, 645–660. Springer.

Isobe, T.; Li, S.; Jia, X.; Yuan, S.; Slabaugh, G.; Xu, C.; Li, Y.-L.; Wang, S.; and Tian, Q. 2020b. Video super-resolution with temporal group attention. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8008–8017.

Jo, Y.; Oh, S. W.; Kang, J.; and Kim, S. J. 2018. Deep video super-resolution network using dynamic upsampling filters without explicit motion compensation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 3224–3232.

Kasten, Y.; Ofri, D.; Wang, O.; and Dekel, T. 2021. Layered neural atlases for consistent video editing. *ACM Transactions on Graphics (TOG)*, 40(6): 1–12.

Keys, R. 1981. Cubic convolution interpolation for digital image processing. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 29(6): 1153–1160.

Kim, J.; Lee, J. K.; and Lee, K. M. 2016. Accurate image super-resolution using very deep convolutional networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 1646–1654.

Ledig, C.; Theis, L.; Huszár, F.; Caballero, J.; Cunningham, A.; Acosta, A.; Aitken, A.; Tejani, A.; Totz, J.; Wang, Z.; et al. 2017. Photo-realistic single image super-resolution using a generative adversarial network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 4681–4690.

Lee, J.; Choi, K. P.; and Jin, K. H. 2022. Learning Local Implicit Fourier Representation for Image Warping. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XVIII*, 182–200. Springer.

Lee, J.; and Jin, K. H. 2022. Local texture estimator for implicit representation function. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1929–1938.

Li, C.; Zhou, A.; and Yao, A. 2022. Omni-Dimensional Dynamic Convolution. In *International Conference on Learning Representations*.

Li, W.; Tao, X.; Guo, T.; Qi, L.; Lu, J.; and Jia, J. 2020. Mucan: Multi-correspondence aggregation network for video super-resolution. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part X 16*, 335–351. Springer.

Liang, J.; Cao, J.; Sun, G.; Zhang, K.; Van Gool, L.; and Timofte, R. 2021. Swinir: Image restoration using swin transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 1833–1844.

Liao, R.; Tao, X.; Li, R.; Ma, Z.; and Jia, J. 2015. Video super-resolution via deep draft-ensemble learning. In *Proceedings of the IEEE International Conference on Computer Vision*, 531–539.

Lim, B.; Son, S.; Kim, H.; Nah, S.; and Mu Lee, K. 2017. Enhanced deep residual networks for single image super-resolution. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 136–144.

Liu, C.; and Sun, D. 2011. A Bayesian approach to adaptive video super resolution. In *CVPR 2011*, 209–216.

Liu, H.; Li, Z.; Shang, F.; Liu, Y.; Wan, L.; Feng, W.; and Timofte, R. 2023. Arbitrary-scale super-resolution via deep

learning: A comprehensive survey. *Information Fusion*, 102015.

Liu, H.; Ruan, Z.; Zhao, P.; Dong, C.; Shang, F.; Liu, Y.; Yang, L.; and Timofte, R. 2022. Video super-resolution based on deep learning: a comprehensive survey. *Artificial Intelligence Review*, 1–55.

Liu, H.; Zhao, P.; Ruan, Z.; Shang, F.; and Liu, Y. 2021. Large motion video super-resolution with dual subnet and multi-stage communicated upsampling. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, 2127–2135.

Lu, Y.; Wang, Z.; Liu, M.; Wang, H.; and Wang, L. 2023. Learning Spatial-Temporal Implicit Neural Representations for Event-Guided Video Super-Resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1557–1567.

Ma, C.; Yu, P.; Lu, J.; and Zhou, J. 2022. Recovering Realistic Details for Magnification-Arbitrary Image Super-Resolution. *IEEE Transactions on Image Processing*, 31: 3669–3683.

Ma, N.; Zhang, X.; Huang, J.; and Sun, J. 2020. Weightnet: Revisiting the design space of weight networks. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XV*, 776–792. Springer.

Mao, X.; Shen, C.; and Yang, Y.-B. 2016. Image restoration using very deep convolutional encoder-decoder networks with symmetric skip connections. *Advances in Neural Information Processing Systems*, 29.

Mildenhall, B.; Srinivasan, P. P.; Tancik, M.; Barron, J. T.; Ramamoorthi, R.; and Ng, R. 2021. Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65(1): 99–106.

Shi, W.; Caballero, J.; Huszár, F.; Totz, J.; Aitken, A. P.; Bishop, R.; Rueckert, D.; and Wang, Z. 2016. Real-time single image and video super-resolution using an efficient subpixel convolutional neural network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 1874–1883.

Siarohin, A.; Lathuilière, S.; Tulyakov, S.; Ricci, E.; and Sebe, N. 2019. Animating arbitrary objects via deep motion transfer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2377–2386.

Son, S.; and Lee, K. M. 2021. SRWarp: Generalized image super-resolution under arbitrary transformation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7782–7791.

Tian, Y.; Zhang, Y.; Fu, Y.; and Xu, C. 2020. Tdan: Temporally-deformable alignment network for video super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3360–3369.

Wang, L.; Guo, Y.; Liu, L.; Lin, Z.; Deng, X.; and An, W. 2020. Deep video super-resolution using HR optical flow estimation. *IEEE Transactions on Image Processing*, 29: 4323–4336.

Wang, L.; Wang, Y.; Lin, Z.; Yang, J.; An, W.; and Guo, Y. 2021. Learning a single network for scale-arbitrary super-resolution. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 4801–4810.

Wang, X.; Chan, K. C.; Yu, K.; Dong, C.; and Change Loy, C. 2019. Edvr: Video restoration with enhanced deformable convolutional networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 0–0.

Wen, W.; Ren, W.; Shi, Y.; Nie, Y.; Zhang, J.; and Cao, X. 2022. Video super-resolution via a spatio-temporal alignment network. *IEEE Transactions on Image Processing*, 31: 1761–1773.

Xue, T.; Chen, B.; Wu, J.; Wei, D.; and Freeman, W. T. 2019. Video enhancement with task-oriented flow. *International Journal of Computer Vision*, 127: 1106–1125.

Yang, B.; Bender, G.; Le, Q. V.; and Ngiam, J. 2019. CondConv: Conditionally Parameterized Convolutions for Efficient Inference. In Wallach, H.; Larochelle, H.; Beygelzimer, A.; d'Alché-Buc, F.; Fox, E.; and Garnett, R., eds., *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Yang, J.; Shen, S.; Yue, H.; and Li, K. 2021. Implicit Transformer Network for Screen Content Image Continuous Super-Resolution. In Ranzato, M.; Beygelzimer, A.; Dauphin, Y.; Liang, P.; and Vaughan, J. W., eds., *Advances in Neural Information Processing Systems*, volume 34, 13304–13315. Curran Associates, Inc.

Ye, X.; and Bilodeau, G.-A. 2022. Continuous conditional video synthesis by neural processes. *arXiv preprint arXiv:2210.05810*.

Yi, P.; Wang, Z.; Jiang, K.; Jiang, J.; Lu, T.; Tian, X.; and Ma, J. 2021. Omniscient video super-resolution. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 4429–4438.

Yi, P.; Wang, Z.; Jiang, K.; Jiang, J.; and Ma, J. 2019. Progressive fusion video super-resolution network via exploiting non-local spatio-temporal correlations. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 3106–3115.

Yu, J.; Liu, J.; Bo, L.; and Mei, T. 2022. Memory-augmented non-local attention for video super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 17834–17843.

Zhang, K.; Gool, L. V.; and Timofte, R. 2020. Deep unfolding network for image super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3217–3226.

Zhang, Y.; Li, K.; Li, K.; Wang, L.; Zhong, B.; and Fu, Y. 2018a. Image super-resolution using very deep residual channel attention networks. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 286–301.

Zhang, Y.; Tian, Y.; Kong, Y.; Zhong, B.; and Fu, Y. 2018b. Residual dense network for image super-resolution. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2472–2481.