Blind Video Super-Resolution based on Implicit Kernels

Qiang Zhu[1], Yuxuan Jiang[2], Shuyuan Zhu[1]*, Fan Zhang[2], David Bull[2], Bing Zeng[1]
[1]University of Electronic Science and Technology of China, [2]University of Bristol
{zhuqiang@std.,eezsy@,eezeng@}uestc.edu.cn,{yuxuan.jiang,fan.zhang,dave.bull}@bristol.ac.uk

Abstract

*Blind video super-resolution (BVSR) is a low-level vision task which aims to generate high-resolution videos from low-resolution counterparts in unknown degradation scenarios. Existing approaches typically predict blur kernels that are spatially invariant in each video frame or even the entire video. These methods do not consider potential spatio-temporal varying degradations in videos, resulting in suboptimal BVSR performance. In this context, we propose a novel **BVSR** model based on **I**mplicit **K**ernels, **BVSR-IK**, which constructs a multi-scale kernel dictionary parameterized by implicit neural representations. It also employs a newly designed recurrent Transformer to predict the coefficient weights for accurate filtering in both frame correction and feature alignment. Experimental results have demonstrated the effectiveness of the proposed BVSR-IK, when compared with four state-of-the-art BVSR models on three commonly used datasets, with BVSR-IK outperforming the second best approach, FMA-Net, by up to 0.59 dB in PSNR. Source code will be available at https://github.com/QZ1-boy/BVSR-IK.*

1. Introduction

Aiming to generate a high-resolution (HR) video from its low-resolution (LR) version, video super-resolution (VSR) has been widely applied in many scenarios, including video streaming [1, 20, 28], surveillance [12, 59], and remote sensing [36, 53]. It is noted that, in some practical cases, LR videos are potentially contaminated by unknown degradations, which makes the task even more challenging. On the contrary, existing generic VSR methods [4, 5, 56, 67, 69, 70] are typically based on known blur kernels [4, 5] (e.g., Gaussian blur kernel) without specifically modeling them during the restoration process. As a result, these VSR methods are not able to effectively capture the intrinsic characteristics of degraded videos [11] in those challenging cases, leading to suboptimal VSR performance.

*Corresponding author

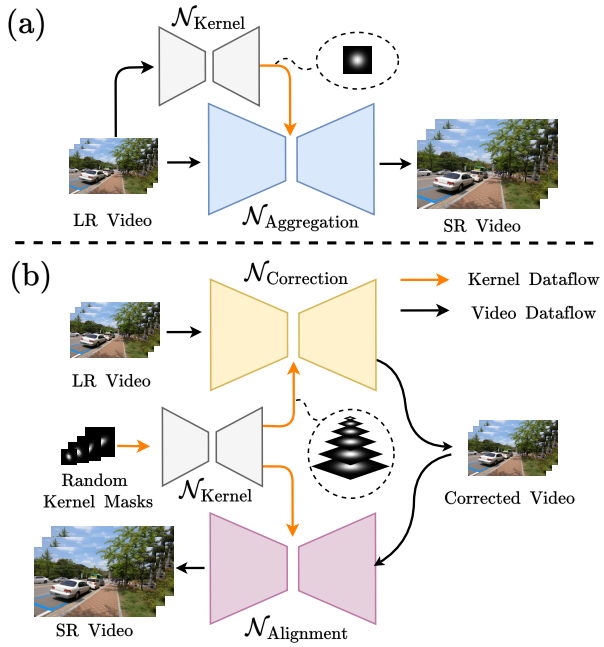

Figure 1. Comparison between existing BVSR methods [2, 40, 54, 60] and our BVSR-IK method. (a) Previous methods predict a single kernel and utilize it to perform super-resolution in an **Aggregation** manner. (b) BVSR-IK (ours) generates an INR-based multi-scale kernel dictionary and applies it in a **Correction-Alignment** manner for BVSR.

To address this issue, different blind VSR (BVSR) methods [2, 18, 26, 40, 49, 54, 61] have been proposed, which are designed to predict specific blur kernels or extract degradations in each inference case. Although these BVSR methods can offer promising super-resolution performance in unknown degradation scenarios, they are also associated with various limitations: (i) these approaches typically assume blur kernels are spatially invariant within each video frame or the whole video, while such an assumption is not applicable in many real cases where blur kernels are spatially variant due to motion and/or focus blurs; (ii) the temporal alignment modules within many BVSR methods [40, 54, 61] assume that video frames share the same degradations [40, 54] - this ignores the temporal inconsistency of degradations

among frames. In both cases, visual artifacts often exhibit (e.g. degradation and misalignment artifacts), in particular, when the degradations are complex.

In this context, this paper proposes **BVSR-IK**, a novel **BVSR** framework (as illustrated in Figure 1) based on **I**mplicit **K**ernels, which generates a multi-scale kernel dictionary using implicit neural representations (INR) to efficiently parameterize multi-scale kernel atoms and implicitly model spatially varying degradations. Moreover, based on the implicit kernels, a new recurrent Transformer is designed to predict coefficient weights by capturing both short-term and long-term scale-aware information. These predicted coefficient weights are then employed for accurate filtering in two modules developed in this framework, Implicit Spatial Correction (ISC) and Implicit Temporal Alignment (ITA), which produce the final super-resolved video. The main contributions are summarized as follows:

- We proposed **the first BVSR model based on implicit kernels**, which predicts multi-scale kernel representations for spatially varying degradations.
- It is also the first time to employ a (modified) **recurrent Transformer for coefficient weight prediction**, which can capture sufficient spatial and temporal information.
- **A new correction-alignment framework** is built for BVSR, based on the predicted implicit multi-scale kernels. Existing methods typically adopt an aggregation-based approach (as shown in Figure 1).

The proposed method has been extensively evaluated on three datasets, compared to four state-of-the-art (SOTA) BVSR methods. The results show that BVSR-IK has achieved consistent performance improvement over FMA-Net [60] (the second best model in the experiment), with up to a 0.59 dB gain in PSNR.

2. Related Work

Blind Image Super-Resolution. Single image super-resolution [7, 10, 19, 46, 64, 68] can offer satisfactory reconstruction performance when blur kernel is known, but perform poorly in unknown degradation scenarios. To address this problem, blind image super-resolution (BISR) methods have been proposed, and can be classified into two categories, kernel prediction (KP) based, and degradation extraction (DE) based. The former typically predicts blur kernels which are further used to guide the SR process. Some advanced techniques have been adopted for kernel estimation, including FKP [27] based on a normalizing flow-based kernel prior and MetaKernelGAN [25] through meta-learning enhanced kernel estimation adaptation. The DE-based methods [32, 41, 48, 52] learn a degradation representation from the degraded image, with notable examples such as DASR [48], CDFormer [33] and DiffTSR [65].

Blind Video Super-Resolution. Similar challenges are also associated with generic video super-resolution models [4, 5, 13, 17, 50, 57], which are based on known degradation conditions. Blind video super-resolution (BVSR) approaches have therefore been proposed [2, 18, 26, 40, 49, 54, 61] based on different network structures and learning methodologies. DynaVSR [26] is one of the earliest works, which employs meta-learning to estimate a downscaled model and adapt to the current input, while DBVSR [40] simultaneously estimates blur kernels, optical flows, and latent images for video restoration. More recently, BSVSR [54] has been proposed based on a multi-scale deformable convolution and deformable attention to enhance spatial details in a coarse-to-fine manner. Finally, FMA-Net [60] is one of the latest contributions that employs 3D convolution layers for degradation learning, which has been reported to offer improved performance over other existing methods.

Implicit Neural Representation (INR) has emerged recently as a new compelling technique to effectively represent visual signals via coordinate-based multi-layer perceptrons (MLPs) [44]. It has been exploited in many application scenarios including 3D reconstruction [3, 9, 35], image/video compression [14, 22, 23, 45], visual signal generation [39, 55], and continuous image super-resolution [6, 8]. However, its application has not been investigated in the field of blind video super-resolution for kernel prediction.

3. Proposed Method

Let $\boldsymbol{I}^{LR} = \{I^{LR}_{t-M:t+M}\} \in \mathbb{R}^{(2M+1)\times H\times W\times 3}$ be a group of degraded LR video frames, the goal of BVSR is to generate its clear SR version $\boldsymbol{I}^{SR} = \{I^{SR}_{t-M:t+M}\} \in \mathbb{R}^{(2M+1)\times sH\times sW\times 3}$, where M, H and W denote the temporal radius, height and width of the input frame group, and s is the SR scale factor. As illustrated in Figure 2, the BVSR-IK framework consists of three primary components: Implicit Kernel Generation (IKG), Implicit Spatial Correlation (ISC) and Implicit Temporal Alignment (ITA) modules. Specifically, random kernel masks are first input into the IKG module to build an implicit multi-scale kernel dictionary $\boldsymbol{\mathcal{D}}_t$ for input video frame, I^{LR}_t. They are then fed into the ISC module, which performs spatial correction for I^{LR}_t. Through a few residual blocks (for feature extraction) [63], the resulting feature $\hat{F}_t$ of the corrected frame $\hat{I}_t$ together with those for two neighboring frames (indexed by $t-1$ and $t+1$) are sequentially taken by two ITA modules (also based on the multi-scale kernel dictionary $\boldsymbol{\mathcal{D}}_t$) to achieve inter-frame temporal alignment through bi-directional propagation, obtaining the forward aligned feature $\widetilde{F}^f_t$ and backward aligned feature $\widetilde{F}^b_t$. Finally, an UP-sampler (UP) is used to process the obtained backward aligned feature and produce the SR video frame I^{SR}_t.

3.1. Implicit Kernel Generation

During the restoration process, the input frame I^{LR}_t is processed a spatially varying deconvolution filtering $\mathcal{F}_d$ and a

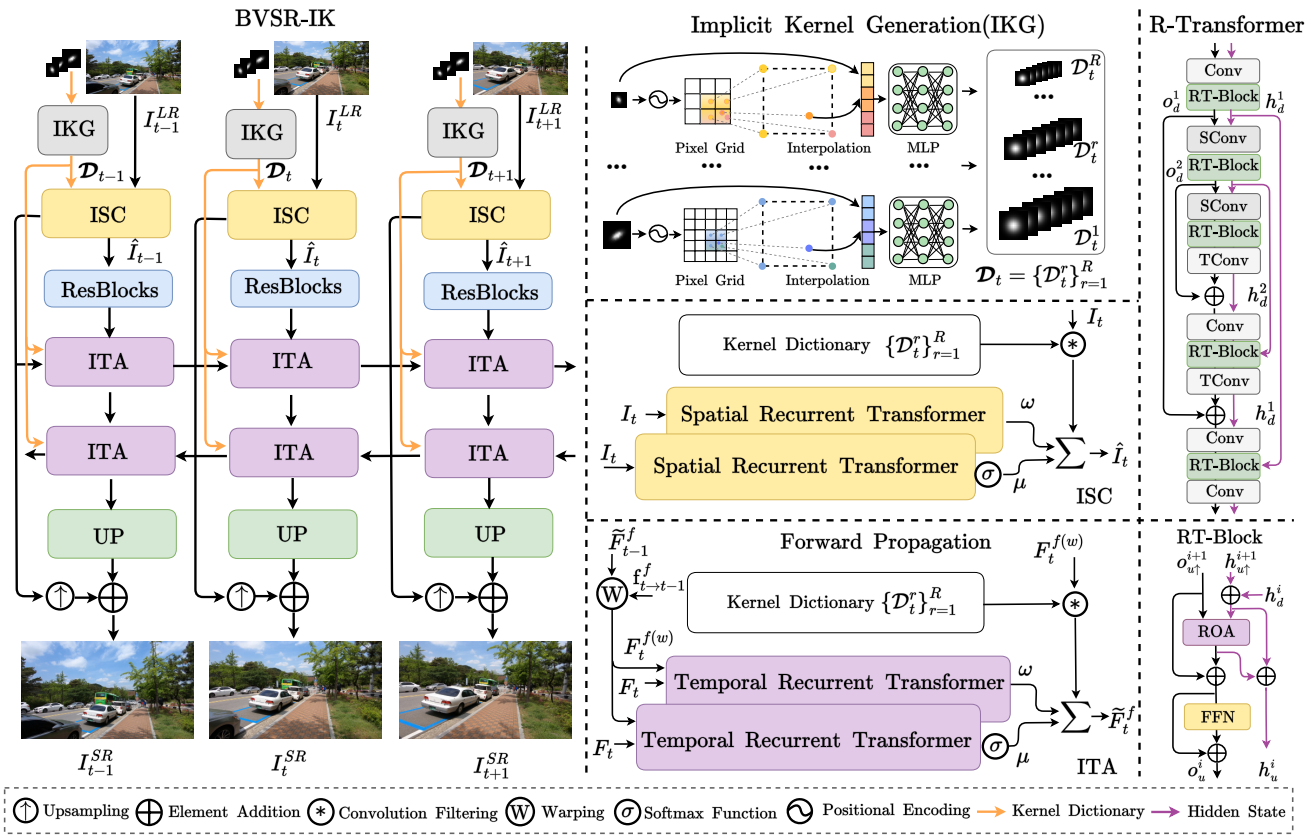

Figure 2. The framework of the BVSR-IK model. The BVSR-IK model consists of three modules, i.e., ISC, ITA, and UP. Each LR video frame is first fed into the ISC module to generate its kernel dictionary and corrected frame. Then, the feature of the corrected frame extracted by residual blocks and the constructed kernel dictionary are fed into the ITA module to achieve the temporal feature alignment through bidirectional propagation. Finally, the aligned feature is fed into the UP module to generate the SR video frame.

set of blur kernels $\mathcal{K} = \{\boldsymbol{k}_{i,j}\}_{i,j}$:

$$\mathcal{F}_d\left(I_t^{LR}; \mathcal{K}\right)[i,j] = \sum_{x,y} \boldsymbol{k}_{i,j}[x,y] I_t^{LR}[i-x, j-y]. \quad (1)$$

Following [15, 42], the blur kernel $\boldsymbol{k} = \sum_{n=1}^{N} \omega_n \boldsymbol{d}_n$ within a low-dimensional space can be expressed as a linear combination of atoms from a dictionary $\mathcal{D} = [\boldsymbol{d}_1, \ldots, \boldsymbol{d}_N]$, where $\omega_n \in \mathbb{R}$ are the coefficient weights.

In this work, to better handle spatially varying degradations and accurately approximate them using the minimum number of atoms, a multi-scale structure is usually introduced to construct the dictionary atoms. Specifically, we define a multi-scale kernel dictionary with R sub-dictionaries, $\{\mathcal{D}^r\}_{r=1}^{R}$, each of which is defined as $\mathcal{D}^r = [\boldsymbol{d}_1^r, \cdots, \boldsymbol{d}_N^r] \subset \mathbb{R}^{M_r \times M_r}$. Here $\boldsymbol{d}_n^r$ denotes the atom at the r-th scale and M_r is the size parameter of the atom, where M_r gets smaller as r increases. The per-pixel blur kernels are then expressed as:

$$\boldsymbol{k}_{i,j} = \sum_{n=1}^{N} \omega_{n,i,j} \boldsymbol{d}_n^{r_{i,j}} \in \mathbb{R}^{M_{r_{i,j}} \times M_{r_{i,j}}}, \quad (2)$$

where $r_{i,j}$ denotes the scale index related to $\boldsymbol{k}_{i,j}$. Note that the atoms within the same sub-dictionary have a uniform size, while their sizes vary across different sub-dictionaries.

Based on the linear representation in Equation 2 and the fact that $\boldsymbol{d}_n^{r_{i,j,j}} * I_t^{LR} = \sum_r \delta\left(r - r_{i,j}\right) \boldsymbol{d}_n^r * I_t^{LR}$ with a Dirac delta function δ, Equation 1 is rewritten as:

$$\begin{aligned} \mathcal{F}_d(I_t^{LR}) &= \sum_{i,j} \mathbf{1}_{i,j} \odot \left(\sum_{n=1}^{N} \omega_{n,i,j} \boldsymbol{d}_n^{r_{i,j}} * I_t^{LR} \right) \\ &= \sum_{i,j} \mathbf{1}_{i,j} \odot \left(\sum_{n=1}^{N} \sum_{r=1}^{R} \mu_{r,i,j} \omega_{n,i,j} \boldsymbol{d}_n^r * I_t^{LR} \right), \end{aligned} \quad (3)$$

in which $\mathbf{1}_{i,j}$ represents an identity matrix, and $\mu_{r,i,j} = \delta\left(r - r_{i,j}\right)$. We can then perform spatially-varying deconvolution filtering by convolving each atom $\boldsymbol{d}_n^r$ in the dictionary with the input I_t^{LR}, and adding the resulting outputs to the spatially-varying weights $\mu_{r,i,j}\omega_{n,i,j}$. The weights $\mu_{r,i,j}$ and $\omega_{n,i,j}$ control the size (adaptive to scene depth) and shape (adaptive to image) of the corresponding blur kernel [42]. The dictionary $\{\mathcal{D}^r\}_{r=1}^{R}$ used in Equation 3 is generated by base atoms $\{\boldsymbol{d}_n\}_{n=1}^{N}$ with the smallest size.

Moreover, to better recover the base atoms, we use implicit neural representation (INR) to re-parameterize the multi-scale kernel dictionary. INR provides a more expressive way to generate multi-scale atoms that cover more frequencies than simple upsampling. Specifically,

$$\boldsymbol{d}_n^r[x,y] = \phi_n(x,y), [x,y] \in [1 \cdots M_r] \times [1 \cdots M_r], \quad (4)$$

where $[x, y]$ is a spatial coordinate, and ϕ is an INR model implemented with a compact MLP. The INR model maps spatial coordinates to a continuous function that implicitly represents the kernel atom. The interpolation process is illustrated in Figure 2. This function can be used to interpolate a kernel atom and create a dictionary of blur kernels with arbitrary scales. Based on this INR-based kernel representation, the learned kernel multi-scale dictionary of the input LR video frame I_t^{LR} can be denoted as $\boldsymbol{\mathcal{D}_t} = \{\mathcal{D}_t^r\}_{r=1}^R$.

3.2. Recurrent Transformer

To explore the scale-aware contextual information from input spatial frames or capture temporal features for generating the accuracy coefficient weights, we design a recurrent Transformer architecture consisting of recurrent Transformer blocks (RT-Block) in a multi-scale structure. This recurrent Transformer has been employed as the backbone to build both the ISC and ITA modules.

As shown in Figure 2, the recurrent Transformer involves operations at three different scales ($d_s = \{1, 2, 3\}$). At each scale level, two inputs, feature branch (black line) and hidden state branch (purple line), are fed into the convolution and RT-Block structure to obtain scale-aware feature and corresponding scale-aware hidden state. The stride convolution (SConv) and transposed convolution (TConv) are applied to downsample and upsample features. In the downsampling stage, RT-Block optimizes the input feature and hidden state and obtains the corresponding scale-aware feature o_d^i, $i \in \{1, 2, 3\}$ and the short-term hidden state h_d^i, $i \in \{1, 2, 3\}$. In the upsampling stage, the previous hidden states h_d^i, $i \in \{1, 2\}$, and the long-term hidden states are passed to the upsampled same resolution RT-Block to achieve the optimization of the current short-term feature o_u^i, $i \in \{1, 2\}$ and the long-term hidden state. The previous features o_d^i, $i \in \{1, 2\}$ are added to o_u^i, $i \in \{1, 2\}$ to enhance the scale-aware feature information. Based on this process, the coefficient weights can be effectively predicted.

In each RT-Block, input feature and hidden state are optimized together based on the new recurrent optimization attention (ROA) and the feed-forward network (FFN) [60]. In the downsampling stage, no long-term hidden state is introduced into the ROA. As shown in Figure 2, the previously upsampled feature $o_{u\uparrow}^{i+1} \in \mathbb{R}^{H \times W \times C}$ and the previously upsampled short-term hidden state $h_{u\uparrow}^{i+1} \in \mathbb{R}^{H \times W \times C}$ are fed into the ROA to interact them for scale-aware information optimization, where H,W,C are the height, width and channel number of the feature. After that, the queries, keys and values are obtained by:

$$Q_o, K_o, V_o = \text{Split}(\mathcal{C}^3(o_{u\uparrow}^{i-1})), \quad (5)$$

$$Q_h, K_h, V_h = \text{Split}(\mathcal{C}^3(h_{u\uparrow}^{i-1})), \quad (6)$$

in which $\mathcal{C}^3(\cdot)$ is a 3×3 convolution and Split is the channel split operation. We then reshape the query and key projections such that their dot-product interaction generates the output feature $A \in \mathbb{R}^{H \times W \times C}$. This is defined as:

$$A = \mathcal{C}^1((V_o + V_h) \cdot \sigma((K_o + K_h) \cdot (Q_o + Q_h)/\alpha)), \quad (7)$$

where $\mathcal{C}^1(\cdot)$ is the 1×1 depth-wise convolution, σ is the Softmax function and $Q_o, O_h \in \mathbb{R}^{HW \times C}, K_o, K_h \in \mathbb{R}^{C \times HW}$. $V_o, V_h \in \mathbb{R}^{HW \times C}$ are obtained after reshaping tensors from the original size $\mathbb{R}^{H \times W \times C}$. α is a learnable scaling parameter [62].

The last convolution of the recurrent Transformer is applied to convert the output feature to the weights $\boldsymbol{\mu}$ and $\boldsymbol{\omega}$. For predicting $\boldsymbol{\mu}$, the Softmax function is used to impose non-negativity and 1-normalization constraints. Such a soft relaxation of weights $\mu_{r,i,j}$ from 0, 1 to [0, 1] can improve the representation accuracy of blur kernels by including the use of additional atoms.

The recurrent Transformer is applied in the ISC module and the ITA module, denoting it as spatial recurrent Transformer (SRT) and temporal recurrent Transformer (TRT), respectively, to explore the spatial scale-aware information and temporal scale-aware information.

3.3. Implicit Spatial Correction

The structure of the ISC module is illustrated in Figure 2. To predict coefficient weights, our designed spatial recurrent Transformers are adopted to predict the coefficient weights $\boldsymbol{\omega} = SRT(I_t)$ and $\boldsymbol{\mu} = \sigma(SRT(I_t))$ for correcting input frames. The input frame is fed into the feature branch and hidden state branch to apply spatial recurrent Transformers.

3.4. Implicit Temporal Alignment

In our model, the previously aligned features, optical flows, and constructed multi-scale kernel dictionary are fed into the ITA module through bidirectional propagation for temporal feature alignment. We present the process of the ITA module in forward propagation, which is illustrated in Figure 2. The backward propagation has the same process. Specifically, the forward aligned feature $\widetilde{F}_{t-1}^f$ is warped by an optical flow $\mathrm{f}_{t\to t-1}^f$ to obtain the forward warped feature $F_t^{f(w)} = \mathcal{W}(\widetilde{F}_{t-1}^f, \mathrm{f}_{t\to t-1}^f)$, where $\mathcal{W}$ is the warping operation. The optical flow is generated by Spynet [43] from corrected frames $\mathrm{f}_{t\to t-1}^f$=Spynet$(\hat{I}_t, \hat{I}_{t-1})$. We then adopt our designed temporal recurrent Transformers to predict the coefficient weights $\boldsymbol{\omega} = \text{TRT}(F_t^{f(w)}, F_t)$ and $\boldsymbol{\mu} = \sigma(\text{TRT}(F_t^{f(w)}, F_t))$ for aligning temporal features.

3.5. Loss Functions

To train the proposed BVSR-IK framework, we combine losses observing the ISC module and the overall performance. The total loss $\mathcal{L}$ is:

$$\begin{aligned}\mathcal{L} &= \mathcal{L}_R + \lambda \mathcal{L}_C \\ &= \mathcal{L}_{Char}(\boldsymbol{I}^{SR}, \boldsymbol{I}^{GT}) + \lambda \mathcal{L}_{Char}(\hat{\boldsymbol{I}}, \boldsymbol{I}^{DN}),\end{aligned} \tag{8}$$

where $\boldsymbol{I}^{GT} = \{I^{GT}_{t-M:t+M}\}$ is the ground-truth HR video, and $\boldsymbol{I}^{DN} = \{I^{DN}_{t-M:t+M}\}$ is the downsampled LR video from $\boldsymbol{I}^{GT}$. $\mathcal{L}_{Char}$ represents the Charbonnier loss [24]. λ is the hyper-parameter.

4. Experimental Setup

Implementation details. In the experiment, we set temporal radius M = 2, scale factor s = 4, the number of residual blocks N_1 = 8 for feature extraction. In the ISC module, we employ R = 7 scales and N = 8 INR-based atoms per scale. We use atoms with sizes 1×1, 3×3, ..., 13×13, where the 1×1 atom is fixed as a delta kernel for modeling in-focus regions. The frequency parameters of the sine activation functions in all INR models are initialized by random sampling from [2, 16]. The UP module consists of N_2 = 13 residual blocks and two pixelshuffle layers for video restoration. During training, Adam optimizer [21] and the Cosine Annealing scheme [34] are employed to optimize the BVSR-IK model. The initial learning rate is set to 1×10^{-4}. We trained our BVSR-IK model for 400 epochs. The training LR patch size is 256 × 256 and the mini-batch size is 7. We set λ to 0.2 in our loss function. The models were implemented using PyTorch and trained with two NVIDIA GeForce RTX 3090 GPUs.

Datasets. Following the previous works [2, 30, 50], we use REDS [38] as the training dataset to train our BVSR-IK model. To evaluate the performance of BVSR methods, three commonly used VSR testing datasets, including REDS4 [38], Vid4 [29] and UDM10 [59], are employed. We implemented two degradation scenarios, i.e., Gaussian blur and realistic motion blur, based on isotropic Gaussian blur kernel and realistic motion blur kernel, respectively. Previous generation of degraded videos [2, 40, 54] only used one kernel for one video, which hardly simulates the degradation change in videos. Therefore, we modified the generation method - one kernel degrades one video frame for more accurate simulation. During training, each HR video frame is blurred and downsampled using a randomly generated kernel for both scenarios. Specifically, for Gaussian blur, the degradation is implemented by applying the randomly generated isotropic Gaussian kernel, where the range of the standard deviation σ (of kernel) is configured as $\sigma \in [0.4, 2.0]$. Moreover, we set the realistic motion blur scenario to simulate the realistic degradation scenario. The degradation is achieved by randomly generating realistic motion blur kernels. The Gaussian blur and realistic motion blur kernels are set to 13 × 13 kernel in terms of size. We use the same approach to generate LR videos in the inference phase.

Evaluation metrics. Aligned with the literature, PSNR, and SSIM [51] have been adopted to evaluate the restoration quality, and tOF [11] is used to measure the temporal consistency of restored videos. Model size, FLOPs, and runtime are also calculated to estimate model complexity.

Benchmark methods. Eight SOTA BISR and BVSR methods have been used to benchmark the performance of the proposed BVSR-IK. These include four BISR models, DASR [48], DIP-DKP [58], DCLS [37] and DARSR [66], and four BVSR approaches, DBVSR [40], BSVSR [54], Self-BVSR [2], FMA-Net [60]. Due to the introduction of a new degradation generation method, we retrained these models based on their open-source implementations to obtain corresponding results.

5. Results and Discussion

5.1. Overall Performance

Quantitative results. Table 1 summarizes the results of the proposed and benchmark methods on three datasets for Gaussian blur and realistic motion blur scenarios. It can be observed that our BVSR-IK model achieves consistent and significant performance improvement over all the other BISR and BVSR models, with a 0.39 dB PSNR gain compared to the second best performer FMA-Net [60], for the Gaussian blur scenario, and 0.59 dB for the realistic motion blur scenario, both on the REDS4 testing dataset.

Qualitative results. Figure 3 provides visual comparison examples for four BVSR methods on three testing datasets under two degradation scenarios. It is evident that the reconstructed results produced by BVSR-IK exhibit better perceptual quality with fewer degradations and finer details compared to other benchmarks.

Model complexity. We evaluate the model complexity of four BVSR methods in terms of FLOPs, runtime, and model size on the REDS4 dataset for the realistic motion blur scenario. Specifically, we implement DBVSR [40], BSVSR [54], Self-BVSR [2], and FMA-Net [60], with M=2 temporal radius to achieve a fair comparison. The FLOPs and runtime are measured for 64×64 LR video frames on a NVIDIA GeForce RTX 3090 GPU for generating HR videos. All the complexity figures are summarized in Table 2. Our BVSR-IK model has the minimal model parameters and competitive complexity. Compared with the current SOTA method, FMA-Net [60], our method has reduced 0.64M parameters and 105.19G FLOPs complexity and obtained a significantly better performance gain.

Scenarios	Types	Methods	REDS4 [38] PSNR↑ / SSIM↑ / tOF↓	Vid4 [29] PSNR↑ / SSIM↑ / tOF↓	UDM10 [59] PSNR↑ / SSIM↑ / tOF↓
Gaussian Blur	Blind SISR	DASR [48]	25.73 / 0.7062 / 4.17	22.98 / 0.6653 / 1.20	27.64 / 0.7661 / 1.30
		DCLS [37]	26.19 / 0.7183 / 3.48	23.28 / 0.6754 / 1.12	28.06 / 0.7804 / 1.08
		DARSR [66]	25.26 / 0.6969 / 3.67	22.25 / 0.6413 / 1.73	27.81 / 0.7689 / 1.17
		DIP-DKP [58]	26.94 / 0.7286 / 3.25	23.50 / 0.6887 / 0.92	28.22 / 0.7898 / 0.96
	Blind VSR	DBVSR [40]	28.50 / 0.8071 / 2.03	23.98 / 0.6747 / 0.72	29.54 / 0.8227 / 0.96
		BSVSR [54]	28.89 / 0.8113 / 1.91	24.15 / 0.6817 / 0.69	29.92 / 0.8423 / 0.89
		Self-BVSR [2]	29.08 / 0.8145 / 2.12	24.28 / 0.6918 / 0.65	30.08 / 0.8661 / 0.85
		FMA-Net [60]	29.11 / 0.8167 / 1.87	24.33 / 0.6925 / 0.68	30.19 / 0.8687 / 0.81
		BVSR-IK (Ours)	**29.50 / 0.8412 / 1.77**	**24.48 / 0.7027 / 0.48**	**30.21 / 0.8692 / 0.77**
Realistic Motion Blur	Blind SISR	DASR [48]	25.49 / 0.7006 / 4.79	22.76 / 0.6133 / 1.43	27.43 / 0.7532 / 1.34
		DCLS [37]	25.93 / 0.7082 / 3.73	22.98 / 0.6247 / 1.29	27.95 / 0.7769 / 1.06
		DARSR [66]	25.10 / 0.6958 / 4.27	22.06 / 0.6043 / 1.66	27.76 / 0.7668 / 1.19
		DIP-DKP [58]	26.75 / 0.7226 / 3.43	23.24 / 0.6488 / 1.09	28.16 / 0.7883 / 0.99
	Blind VSR	DBVSR [40]	28.03 / 0.7882 / 2.83	23.73 / 0.6494 / 0.76	29.47 / 0.8216 / 0.94
		BSVSR [54]	28.58 / 0.8041 / 1.92	23.97 / 0.6559 / 0.69	29.82 / 0.8364 / 0.91
		Self-BVSR [2]	28.34 / 0.7926 / 2.21	24.04 / 0.6735 / 0.66	30.03 / 0.8568 / 0.87
		FMA-Net [60]	28.89 / 0.8094 / 1.89	24.16 / 0.6718 / 0.67	30.12 / 0.8675 / 0.83
		BVSR-IK (Ours)	**29.48 / 0.8303 / 1.82**	**24.52 / 0.7029 / 0.48**	**30.27 / 0.8700 / 0.74**

Table 1. Quantitative evaluations by PSNR (dB), SSIM, and tOF on REDS4 [38], Vid4 [29], and UDM10 [59] datasets for Gaussian blur and realistic motion blur scenarios. The results are tested on the Y-channel. The best results in each case have been marked in **bold.**

Methods	Frame	Param.(M) ↓	FLOPs(G) ↓	Runtime(s) ↓
DBVSR [40]	5	14.10	158.69	0.86
BSVSR [54]	5	20.83	489.16	1.81
Self-BVSR [2]	5	18.20	281.45	1.29
FMA-Net [60]	5	9.94	564.42	2.07
BVSR-IK (Ours)	5	9.30	459.23	1.58

Table 2. Comparisons of parameters and complexity. The FLOPs and Runtime are computed with 64×64 LR video resolution.

5.2. Ablation Study

To further verify the effectiveness of the main contributions in this work, we have designed an ablation study based on the REDS4 dataset for the realistic motion blur scenario. The corresponding results are provided in Table 3.

Effectiveness of the designed modules and loss. We first tested the contribution of our main modules, i.e., ISC module and ITA module, by creating the following variants. (v1.1) w/o ISC - the ISC module was removed from BVSR-IK; (v1.2) w/o ITA - the ITA module was removed from the full model and aligned features were generated by warping with optical flow. We further verified two structures, i.e., recurrent structure in our designed recurrent Transformer and bi-directional propagation in BVSR-IK, resulting in (v1.3) w/o Rec - the hidden state branch was removed from our recurrent Transformer and the corrected frame and warped feature only fed into Transformer; (v1.4) w/o Bi-dir - the ITA module in the backward propagation was removed from our framework; (v1.5) w/o $\mathcal{L}_C$ - the correction loss was removed when training our BVSR-IK model. It can be observed from Table 3 (a) that our ISC module can bring 0.30 dB PSNR gain and only increase 11.75G FLOPs complexity, and the ITA module achieves 0.62 dB PSNR gain with a 85.75G FLOPs increase. Recurrent structure, bidirectional propagation and correction loss can contribute 0.28 dB, 0.19 dB and 0.08 dB PSNR gains, respectively. We also illustrate the visual results of the above variants and our full model in Figure 4. These results present the visual quality improvement when using our designed modules and loss.

Models	PSNR↑ / SSIM↑ / tOF↓	Param.(M)↓	FLOPs(G)↓	R.T.(s)↓
(a) Ablation Study of Modules and Loss				
(v1.1) *w/o* ISC	29.18 / 0.8128 / 1.87	8.12	447.45	1.42
(v1.2) *w/o* ITA	28.86 / 0.8089 / 1.92	5.50	373.51	0.97
(v1.3) *w/o* Rec	29.20 / 0.8139 / 1.86	8.55	435.13	1.48
(v1.4) *w/o* Bi-dir	29.29 / 0.8158 / 1.85	7.40	346.16	0.82
(v1.5) *w/o* $\mathcal{L}_C$	29.40 / 0.8289 / 1.84	9.30	459.23	1.58
(b) Comparison of Different Kernel Prediction (KP) with IK				
(v2.1) [2]'s KP	29.16 / 0.8127 / 1.95	8.71	423.65	1.45
(v2.2) [60]'s KP	29.35 / 0.8278 / 1.88	9.24	484.26	1.78
(v2.3) DKP [16]	28.92 / 0.8093 / 2.03	6.85	394.26	1.27
(c) Comparison of Different Correction Filters with ISC				
(v3.1) Corr Filter [16]	29.11 / 0.8107 / 1.92	8.21	451.21	1.45
(v3.2) Corr Filter [66]	29.20 / 0.8140 / 1.89	8.79	468.92	1.51
(v3.3) MISCFilter [31]	29.35 / 0.8284 / 1.88	9.36	463.10	1.60
(d) Comparison of Different Alignment Modules with ITA				
(v4.1) Flow [4]	28.84 / 0.8082 / 1.98	5.57	373.60	0.99
(v4.2) DCN [47]	29.08 / 0.8106 / 1.91	6.27	394.56	1.23
(v4.3) FGDA [5]	29.45 / 0.8309 / 1.86	9.58	467.32	1.66
(v4.4) FGDF [60]	29.21 / 0.8167 / 1.89	6.12	386.32	1.06
BVSR-IK (Ours)	29.48 / 0.8303 / 1.82	9.30	459.23	1.58

Table 3. Results of the ablation study.

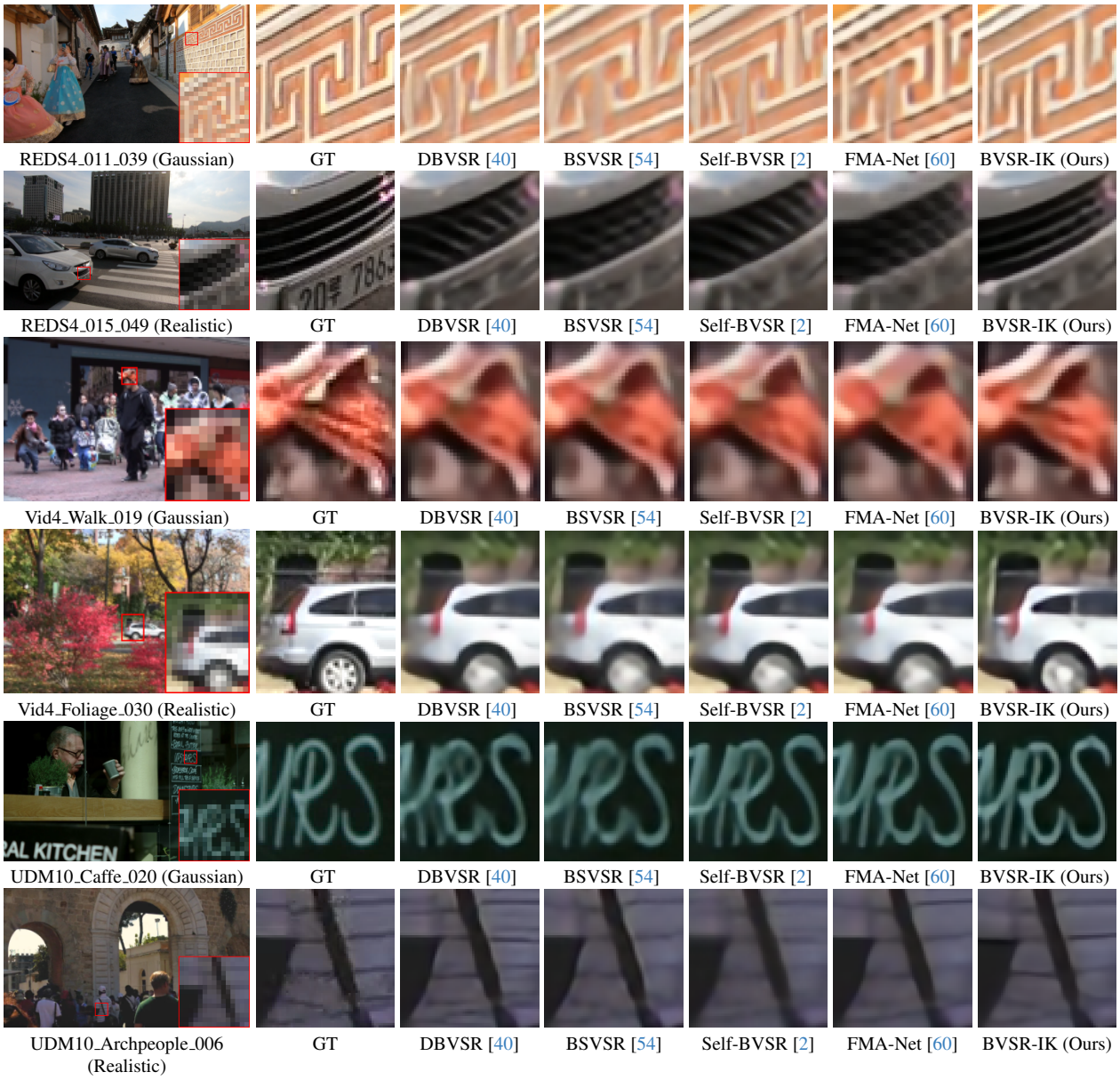

Figure 3. Visual results on REDS4 [38], Vid4 [29] and UDM10 [59] datasets for Gaussian blur and realistic motion blur scenarios.

Comparison with existing kernel predictions. To compare our designed kernel dictionary with existing kernel prediction methods. We implemented two kernel predictions in BVSR methods, i.e., Self-BVSR [2] and FMA-Net [60], and an unsupervised kernel prediction method in BISR method, i.e., DKP [58] to replace our designed kernel dictionary in our BVSR-IK model. The corresponding results were reported in Table 3 (b). Since Self-BVSR [2] and FMA-Net [60] only predicted one kernel for one frame, the diversity of kernel representation was limited. Additionally, DKP [58] improved the kernel representation by a Markov chain Monte Carlo sampling process on random kernel distributions, but did not consider the multi-scale kernel distributions, resulting in insufficient kernel representation and further information exploration.

Comparison with existing correction filters. To compare the existing correction filters in our ISC module. We replaced our ISC module with three representative correct filters [16, 31, 66] in our BVSR-IK model and obtained (v3.1) Corr Filter [16], (v3.2) Corr Filter [66] and (v3.1) MISFilter [31]. The results in Table 3 (c) have confirmed the superior performance of our ISC module over others.

Comparison with existing alignment modules. To compare our designed ITA module with the existing align-

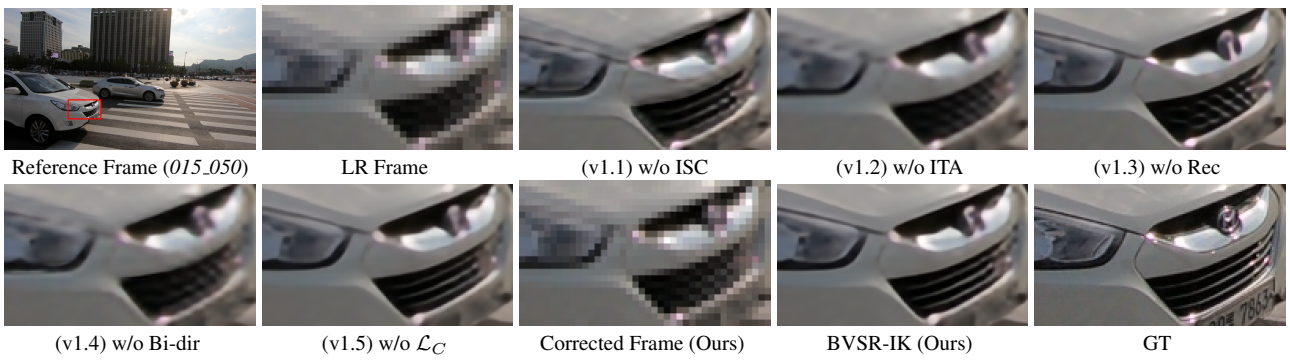

Figure 4. Visualization of ablation studies on BVSR-IK on REDS4 [38] dataset frame for realistic motion blur scenario.

R	PSNR↑ / SSIM↑ / tOF↓	Param.(M)	N	PSNR↑ / SSIM↑ / tOF↓	Param.(M)
4	29.16 / 0.8105 / 1.95	9.10	2	29.21 / 0.8263 / 1.94	6.59
5	29.24 / 0.8303 / 1.92	9.18	4	29.30 / 0.8285 / 1.89	7.45
6	29.37 / 0.8278 / 1.86	9.24	6	29.42 / 0.8291 / 1.86	8.39
7	**29.48 / 0.8303 / 1.82**	9.30	8	29.48 / 0.8303 / 1.82	9.30
8	29.44 / 0.8307 / 1.81	9.37	10	29.50 / 0.8309 / 1.80	10.20
9	29.42 / 0.8291 / 1.83	9.44	12	**29.53 / 0.8318 / 1.78**	11.11

Table 4. Ablation study of the kernel scale R and atom number N.

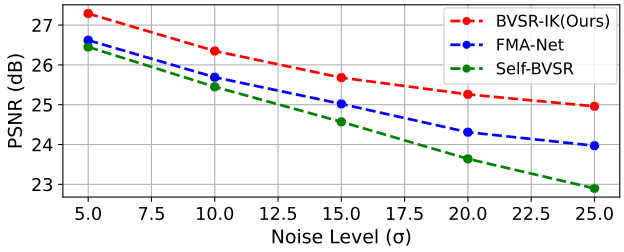

Figure 5. Robustness of BVSR models for noise scenario.

ment module, we used four recent alignments, i.e., flow-based [4], DCN-based [47], FGDA-based [5], and FGDF-based [60] to replace our ITA module in our BVSR-IK model. Four corresponding variant models were created: (v4.1) Flow, (v4.2), DCN (v4.3) FGDA, and (v4.4) FGDF, respectively. The results in Table 3 show that our ITA module has a similar performance to the FGDA module, but with lower complexity.

Effectiveness of kernel scale and atom number. To confirm the kernel dictionary scale R and atom number N in our multi-scale kernel dictionary, we set the different R and N, respectively, to construct the BVSR model. The corresponding results are presented in Table 4. The max kernel size in kernel dictionary is calculated by $k_{max} = 2R - 1$. We found that with R increasing, the BVSR performance was improved. When larger than $R = 8$ (k_{max}=15), the performance was dropped, which may be due to a large kernel causing the extra blur information. In addition, as N increases, the performance improves, but the model complexity also increases. To avoid excessive model complexity and model size, we set $N = 8$ in our work.

Noise	Self-BVSR [2]	FMA-Net [60]	BVSR-IK (Ours)
σ=5	26.45 / 0.6927 / 3.69	26.62 / 0.7052 / 3.62	27.29 / 0.7390 / 3.60
σ=10	25.46 / 0.6125 / 4.89	25.69 / 0.6567 / 4.46	26.35 / 0.7039 / 4.41
σ=15	24.57 / 0.5332 / 4.94	25.02 / 0.5851 / 4.90	25.68 / 0.6782 / 4.89
σ=20	23.64 / 0.4619 / 5.24	24.31 / 0.5300 / 5.21	25.26 / 0.6592 / 4.98
σ=25	22.90 / 0.4031 / 5.46	23.97 / 0.5252 / 5.28	24.96 / 0.6427 / 5.07

Table 5. Verify the robustness of the BVSR models.

Robustness of the BVSR-IK. To confirm the robustness of BVSR models on the noise scenario, we retrain Self-BVSR [2], FMA-Net [60] and our model on the realistic motion blur scenario with noise. The noise is randomly added in each LR blurred video frame, and the range was set to [0, 25]. We added five noise levels, i.e., 5, 10, 15, 20, and 25, on the REDS4 dataset to evaluate the model robustness. The corresponding results are reported in Table 5 and Figure 5. It is found from these results that our BVSR-IK has better robustness for the noise scenario compared with self-BVSR and FMA-Net, when the noise level increases.

6. Conclusion

In this paper, we proposed a novel blind video super-resolution framework, BVSR-IK, which is the first attempt to employ INR to generate a multi-scale kernel dictionary for spatially varying degradations in blind super-resolution. Based on the implicit kernel dictionary, a new recurrent Transformer is designed to predict coefficient weights by capturing both short-term and long-term scale-aware information, which are then employed for accurate filtering for implicit spatial correction and implicit temporal alignment module. Our model was evaluated on commonly used datasets and achieved consistent and evident performance gains over SOTA BVSR methods.

Acknowledgments

This work was supported by the Natural Science Foundation of Sichuan Province under Grant 2023NSFSC1972, the China Scholarship Council, the University of Bristol, and the UKRI MyWorld Strength in Places Programme (SIPF00006/1).

References

[1] Mariana Afonso, Fan Zhang, and David R Bull. Video compression based on spatio-temporal resolution adaptation. *IEEE Transactions on Circuits and Systems for Video Technology*, 29(1):275–280, 2018. 1

[2] Haoran Bai and Jinshan Pan. Self-supervised deep blind video super-resolution. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024. 1, 2, 5, 6, 7, 8

[3] Jonathan T Barron, Ben Mildenhall, Matthew Tancik, Peter Hedman, Ricardo Martin-Brualla, and Pratul P Srinivasan. Mip-nerf: A multiscale representation for anti-aliasing neural radiance fields. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5855–5864, 2021. 2

[4] Kelvin CK Chan, Xintao Wang, Ke Yu, Chao Dong, and Chen Change Loy. 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*, pages 4947–4956, 2021. 1, 2, 6, 8

[5] Kelvin CK Chan, Shangchen Zhou, Xiangyu Xu, and Chen Change Loy. Basicvsr++: Improving video super-resolution with enhanced propagation and alignment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5972–5981, 2022. 1, 2, 6, 8

[6] Hao-Wei Chen, Yu-Syuan Xu, Min-Fong Hong, Yi-Min Tsai, Hsien-Kai Kuo, and Chun-Yi Lee. Cascaded local implicit Transformer for arbitrary-scale super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18257–18267, 2023. 2

[7] Xiangyu Chen, Xintao Wang, Jiantao Zhou, Yu Qiao, and Chao Dong. Activating more pixels in image super-resolution Transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22367–22377, 2023. 2

[8] Yinbo Chen, Sifei Liu, and Xiaolong Wang. Learning continuous image representation with local implicit image function. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8628–8638, 2021. 2

[9] Zhiqin Chen and Hao Zhang. Learning implicit fields for generative shape modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5939–5948, 2019. 2

[10] Zheng Chen, Yulun Zhang, Jinjin Gu, Linghe Kong, Xiaokang Yang, and Fisher Yu. Dual aggregation Transformer for image super-resolution. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12312–12321, 2023. 2

[11] Mengyu Chu, You Xie, Jonas Mayer, Laura Leal-Taixé, and Nils Thuerey. Learning temporal coherence via self-supervision for GAN-based video generation. *ACM Transactions on Graphics*, 39(4):75–1, 2020. 1, 5

[12] Amar B Deshmukh and N Usha Rani. Fractional-grey wolf optimizer-based kernel weighted regression model for multi-view face video super resolution. *International Journal of Machine Learning and Cybernetics*, 10:859–877, 2019. 1

[13] Dario Fuoli, Shuhang Gu, and Radu Timofte. Efficient video super-resolution through recurrent latent space propagation. In *2019 IEEE/CVF International Conference on Computer Vision Workshop*, pages 3476–3485. IEEE, 2019. 2

[14] Ge Gao, Ho Man Kwan, Fan Zhang, and David Bull. Pnvc: Towards practical inr-based video compression. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 3068–3076, 2025. 2

[15] Shaojie Guo, Haofei Song, Qingli Li, and Yan Wang. Spatially-variant degradation model for dataset-free super-resolution. In *European Conference on Computer Vision*, pages 340–357. Springer, 2024. 3

[16] Shady Abu Hussein, Tom Tirer, and Raja Giryes. Correction filter for single image super-resolution: Robustifying off-the-shelf deep super-resolvers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1428–1437, 2020. 6, 7

[17] Takashi Isobe, Xu Jia, Xin Tao, Changlin Li, Ruihuang Li, Yongjie Shi, Jing Mu, Huchuan Lu, and Yu-Wing Tai. Look back and forth: Video super-resolution with explicit temporal difference modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17411–17420, 2022. 2

[18] Mehran Jeelani, Noshaba Cheema, Klaus Illgner-Fehns, Philipp Slusallek, Sunil Jaiswal, et al. Expanding synthetic real-world degradations for blind video super resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1199–1208, 2023. 1, 2

[19] Yuxuan Jiang, Chen Feng, Fan Zhang, and David Bull. Mtkd: Multi-teacher knowledge distillation for image super-resolution. In *European Conference on Computer Vision*, pages 364–382. Springer, 2024. 2

[20] Jung-Heum Kang, Muhammad Salman Ali, Hye-Won Jeong, Chang-Kyun Choi, Younhee Kim, Se Yoon Jeong, Sung-Ho Bae, and Hui Yong Kim. A super-resolution-based feature map compression for machine-oriented video coding. *IEEE Access*, 11:34198–34209, 2023. 1

[21] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 5

[22] Ho Man Kwan, Ge Gao, Fan Zhang, Andrew Gower, and David Bull. Hinerv: Video compression with hierarchical encoding-based neural representation. *Advances in Neural Information Processing Systems*, 36:72692–72704, 2023. 2

[23] Ho Man Kwan, Ge Gao, Fan Zhang, Andrew Gower, and David Bull. Nvrc: Neural video representation compression. *Advances in Neural Information Processing Systems*, 37:132440–132462, 2024. 2

[24] Wei-Sheng Lai, Jia-Bin Huang, Narendra Ahuja, and Ming-Hsuan Yang. Fast and accurate image super-resolution with deep laplacian pyramid networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(11):2599–2613, 2018. 5

[25] Royson Lee, Rui Li, Stylianos Venieris, Timothy Hospedales, Ferenc Huszár, and Nicholas D Lane. Meta-learned kernel for blind super-resolution kernel estimation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1496–1505, 2024. 2

[26] Suyoung Lee, Myungsub Choi, and Kyoung Mu Lee. Dynavsr: Dynamic adaptive blind video super-resolution. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2093–2102, 2021. 1, 2

[27] Jingyun Liang, Kai Zhang, Shuhang Gu, Luc Van Gool, and Radu Timofte. Flow-based kernel prior with application to blind super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10601–10610, 2021. 2

[28] Chaoyi Lin, Yue Li, Junru Li, Kai Zhang, and Li Zhang. Luma-only resampling-based video coding with CNN-based super resolution. In *2023 IEEE International Conference on Visual Communications and Image Processing*, pages 1–5, 2023. 1

[29] Ce Liu and Deqing Sun. On bayesian adaptive video super resolution. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(2):346–360, 2013. 5, 6, 7

[30] Chengxu Liu, Huan Yang, Jianlong Fu, and Xueming Qian. Learning trajectory-aware Transformer for video super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5687–5696, 2022. 5

[31] Chengxu Liu, Xuan Wang, Xiangyu Xu, Ruhao Tian, Shuai Li, Xueming Qian, and Ming-Hsuan Yang. Motion-adaptive separable collaborative filters for blind motion deblurring. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 25595–25605, 2024. 6, 7

[32] Qingguo Liu, Pan Gao, Kang Han, Ningzhong Liu, and Wei Xiang. Degradation-aware self-attention based Transformer for blind image super-resolution. *IEEE Transactions on Multimedia*, 2024. 2

[33] Qingguo Liu, Chenyi Zhuang, Pan Gao, and Jie Qin. Cd-former: When degradation prediction embraces diffusion model for blind image super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7455–7464, 2024. 2

[34] Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts. *arXiv preprint arXiv:1608.03983*, 2016. 5

[35] Yujie Lu, Long Wan, Nayu Ding, Yulong Wang, Shuhan Shen, Shen Cai, and Lin Gao. Unsigned orthogonal distance fields: An accurate neural implicit representation for diverse 3d shapes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20551–20560, 2024. 2

[36] Yimin Luo, Liguo Zhou, Shu Wang, and Zhongyuan Wang. Video satellite imagery super resolution via convolutional neural networks. *IEEE Geoscience and Remote Sensing Letters*, 14(12):2398–2402, 2017. 1

[37] Ziwei Luo, Haibin Huang, Lei Yu, Youwei Li, Haoqiang Fan, and Shuaicheng Liu. Deep constrained least squares for blind image super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17642–17652, 2022. 5, 6

[38] Seungjun Nah, Sungyong Baik, Seokil Hong, Gyeongsik Moon, Sanghyun Son, Radu Timofte, and Kyoung Mu Lee. Ntire 2019 challenge on video deblurring and super-resolution: Dataset and study. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 0–0, 2019. 5, 6, 7, 8

[39] Seonghyeon Nam, Marcus A Brubaker, and Michael S Brown. Neural image representations for multi-image fusion and layer separation. In *European conference on computer vision*, pages 216–232. Springer, 2022. 2

[40] Jinshan Pan, Haoran Bai, Jiangxin Dong, Jiawei Zhang, and Jinhui Tang. Deep blind video super-resolution. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4811–4820, 2021. 1, 2, 5, 6, 7

[41] Yajun Qiu, Qiang Zhu, Shuyuan Zhu, and Bing Zeng. Dual circle contrastive learning-based blind image super-resolution. *IEEE Transactions on Circuits and Systems for Video Technology*, 2023. 2

[42] Yuhui Quan, Xin Yao, and Hui Ji. Single image defocus deblurring via implicit neural inverse kernels. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12600–12610, 2023. 3

[43] Anurag Ranjan and Michael J Black. Optical flow estimation using a spatial pyramid network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4161–4170, 2017. 4

[44] Vishwanath Saragadam, Jasper Tan, Guha Balakrishnan, Richard G Baraniuk, and Ashok Veeraraghavan. Miner: Multiscale implicit neural representation. In *European Conference on Computer Vision*, pages 318–333. Springer, 2022. 2

[45] Yannick Strümpler, Janis Postels, Ren Yang, Luc Van Gool, and Federico Tombari. Implicit neural representations for image compression. In *European Conference on Computer Vision*, pages 74–91. Springer, 2022. 2

[46] Long Sun, Jiangxin Dong, Jinhui Tang, and Jinshan Pan. Spatially-adaptive feature modulation for efficient image super-resolution. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13190–13199, 2023. 2

[47] Yapeng Tian, Yulun Zhang, Yun Fu, and Chenliang Xu. TDAN: Temporally-deformable alignment network for video super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3360–3369, 2020. 6, 8

[48] Longguang Wang, Yingqian Wang, Xiaoyu Dong, Qingyu Xu, Jungang Yang, Wei An, and Yulan Guo. Unsupervised degradation representation learning for blind super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10581–10590, 2021. 2, 5, 6

[49] Ruohao Wang, Xiaohui Liu, Zhilu Zhang, Xiaohe Wu, Chun-Mei Feng, Lei Zhang, and Wangmeng Zuo. Benchmark dataset and effective inter-frame alignment for real-world video super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1168–1177, 2023. 1, 2

[50] Xintao Wang, Kelvin CK Chan, Ke Yu, Chao Dong, and Chen Change Loy. EDVR: Video restoration with enhanced deformable convolutional networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 1954–1963, 2019. 2, 5

[51] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, 2004. 5

[52] Bin Xia, Yulun Zhang, Yitong Wang, Yapeng Tian, Wenming Yang, Radu Timofte, and Luc Van Gool. Knowledge distillation based degradation estimation for blind super-resolution. *arXiv preprint arXiv:2211.16928*, 2022. 2

[53] Yi Xiao, Qiangqiang Yuan, Kui Jiang, Xianyu Jin, Jiang He, Liangpei Zhang, and Chia-wen Lin. Local-global temporal difference learning for satellite video super-resolution. *IEEE Transactions on Circuits and Systems for Video Technology*, 2023. 1

[54] Yi Xiao, Qiangqiang Yuan, Qiang Zhang, and Liangpei Zhang. Deep blind super-resolution for satellite video. *IEEE Transactions on Geoscience and Remote Sensing*, 2023. 1, 2, 5, 6, 7

[55] Dejia Xu, Peihao Wang, Yifan Jiang, Zhiwen Fan, and Zhangyang Wang. Signal processing for implicit neural representations. *Advances in Neural Information Processing Systems*, 35:13404–13418, 2022. 2

[56] Kai Xu, Ziwei Yu, Xin Wang, Michael Bi Mi, and Angela Yao. Enhancing video super-resolution via implicit resampling-based alignment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2546–2555, 2024. 1

[57] Tianfan Xue, Baian Chen, Jiajun Wu, Donglai Wei, and William T Freeman. Video enhancement with task-oriented flow. *International Journal of Computer Vision*, 127:1106–1125, 2019. 2

[58] Zhixiong Yang, Jingyuan Xia, Shengxi Li, Xinghua Huang, Shuanghui Zhang, Zhen Liu, Yaowen Fu, and Yongxiang Liu. A dynamic kernel prior model for unsupervised blind image super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26046–26056, 2024. 5, 6, 7

[59] Peng Yi, Zhongyuan Wang, Kui Jiang, Junjun Jiang, and Jiayi Ma. Progressive fusion video super-resolution network via exploiting non-local spatio-temporal correlations. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3106–3115, 2019. 1, 5, 6, 7

[60] Geunhyuk Youk, Jihyong Oh, and Munchurl Kim. Fmanet: Flow-guided dynamic filtering and iterative feature refinement with multi-attention for joint video super-resolution and deblurring. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 44–55, 2024. 1, 2, 4, 5, 6, 7, 8

[61] Jun-Seok Yun, Min Hyuk Kim, Hyung-Il Kim, and Seok Bong Yoo. Kernel adaptive memory network for blind video super-resolution. *Expert Systems with Applications*, 238:122252, 2024. 1, 2

[62] Syed Waqas Zamir, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, and Ming-Hsuan Yang. Restormer: Efficient Transformer for high-resolution image restoration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5728–5739, 2022. 4

[63] Yulun Zhang, Kunpeng Li, Kai Li, Lichen Wang, Bineng Zhong, and Yun Fu. Image super-resolution using very deep residual channel attention networks. In *Proceedings of the European Conference on Computer Vision*, pages 286–301, 2018. 2

[64] Yulun Zhang, Yapeng Tian, Yu Kong, Bineng Zhong, and Yun Fu. Residual dense network for image super-resolution. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2472–2481, 2018. 2

[65] Yuzhe Zhang, Jiawei Zhang, Hao Li, Zhouxia Wang, Luwei Hou, Dongqing Zou, and Liheng Bian. Diffusion-based blind text image super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 25827–25836, 2024. 2

[66] Hongyang Zhou, Xiaobin Zhu, Jianqing Zhu, Zheng Han, Shi-Xue Zhang, Jingyan Qin, and Xu-Cheng Yin. Learning correction filter via degradation-adaptive regression for blind single image super-resolution. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12365–12375, 2023. 5, 6, 7

[67] Xingyu Zhou, Leheng Zhang, Xiaorui Zhao, Keze Wang, Leida Li, and Shuhang Gu. Video super-resolution Transformer with masked inter&intra-frame attention. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 25399–25408, 2024. 1

[68] Qiang Zhu, Pengfei Li, and Qianhui Li. Attention retractable frequency fusion Transformer for image super resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1756–1763, 2023. 2

[69] Qiang Zhu, Feiyu Chen, Yu Liu, Shuyuan Zhu, and Bing Zeng. Deep compressed video super-resolution with guidance of coding priors. *IEEE Transactions on Broadcasting*, 2024. 1

[70] Qiang Zhu, Feiyu Chen, Shuyuan Zhu, Yu Liu, Xue Zhou, Ruiqin Xiong, and Bing Zeng. Dvsrnet: Deep video super-resolution based on progressive deformable alignment and temporal-sparse enhancement. *IEEE Transactions on Neural Networks and Learning Systems*, 2024. 1