Multi-Frequency Representation Enhancement with Privilege Information for Video Super-Resolution

Fei Li [*]
China Agricultural University
leefly072@126.com

Linfeng Zhang [*]
Tsinghua University
zhanglinfeng1997@outlook.com

Zikun Liu
Samsung Research China
zikun.liu@samsung.com

Juan Lei
Samsung Research China
juan.lei@samsung.com

Zhenbo Li ✉
China Agricultural University
lizb@cau.edu.cn

Abstract

CNN's limited receptive field restricts its ability to capture long-range spatial-temporal dependencies, leading to unsatisfactory performance in video super-resolution (VSR). To tackle this challenge, this paper presents a novel multi-frequency representation enhancement module (MFE) that performs spatial-temporal information aggregation in the frequency domain. Specifically, MFE mainly includes a spatial-frequency representation enhancement branch which captures the long-range dependency in the spatial dimension, and an energy frequency representation enhancement branch to obtain the inter-channel feature relationship. Moreover, a novel model training method named privilege training is proposed to encode the privilege information from high-resolution videos to facilitate model training. With these two methods, we introduce a new VSR model named MFPI, which outperforms state-of-the-art methods by a large margin while maintaining good efficiency on various datasets, including REDS4, Vimeo, Vid4, and UDM10.

1. Introduction

Video super-resolution (VSR), which restores high-resolution (HR) video frames from their highly related but unaligned low-resolution (LR) video frames, is well-desired in various real-world applications [11, 2]. Compared with single image super-resolution (SISR), VSR is much more challenging as it aggregates information from multiple related but misaligned frames in the input video. Hence, an ideal VSR model is expected to leverage the spatial information from single images but also integrate the temporal semantic features from multiple frames [45, 50, 18, 3].

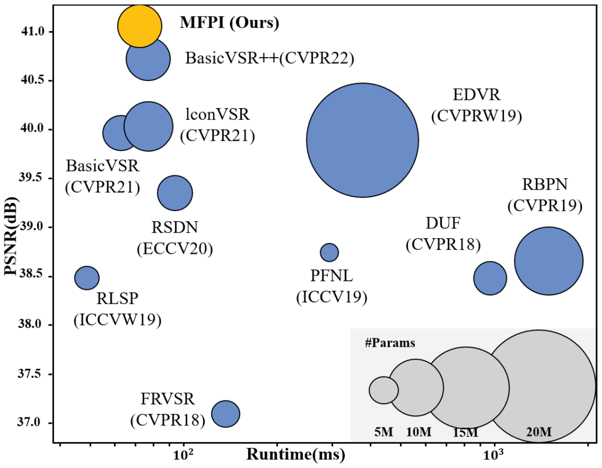


Figure 1: VSR performance comparison on UDM10 [65] in terms of PSNR (dB), runtime (ms), and parameters (M). MFPI outperforms SOTA methods with high efficiency.

CNN-based models have become the standard choice for VSR due to their simplicity and efficiency. While recent advances in CNN-based VSR, such as residual learning [2, 73], dense connections[59], hierarchical structures [20], and multi-scale frameworks [63, 34, 48], have shown impressive performance, some potential problems still exist. Firstly, CNN's limited receptive field hinders it from modeling long-range spatial dependencies, making it unable to capture complex semantic information. Secondly, convolutional filters have fixed scales and weights, limiting their performance on large motions and adaptability to diverse inputs. In comparison, with the advantages of self-attention layers, ViTs are effective in processing inputs with long-term spatial dependencies. However, the stack

[*] The first two authors contribute equally. This work is done during the internship of F. Li in Samsung. ✉ Z. Li is the corresponding author.

of self-attention layers is accompanied with a large computation budget, which results in inferior inference efficiency on edge devices [29]; A natural question arises: *can CNNs capture long-range dependencies like ViTs?*

In this paper, we propose to answer this question from a frequency perspective. Thanks to their ability to capture global representation, frequency-based models have been widely utilized in single image processing [72, 41, 71]. Unfortunately, their performance in video super-resolution (VSR) is usually limited due to several reasons. (1) Compared with single images, videos are composed of multiple related but misaligned images, which contain both spatial and temporal information. However, traditional frequency methods have *fixed paradigms and coefficients* [56, 66, 40], making them not able to capture the complex information in videos. (2) Traditional frequency methods are sensitive to the noise in images [64, 14]. This disadvantage becomes more fatal in videos where the noise of different frames can be accumulated.

To address these challenges, we propose a novel module named multi-frequency representation enhancement (MFE), which aggregates information in the frequency domain. Unlike convolutional and self-attention layers that operate in the feature space, MFE directly manipulates the energy feature map in the frequency domain through three branches: 1) a spatial-frequency representation enhancement branch (SFE) that utilizes a Fast Fourier Transform (FFT) and large kernel convolution layer to capture arbitrary interactions among spatial and long-range dependent features, 2) an energy-frequency representation enhancement branch (EFE) that employs a novel energy discrete cosine transform (DCT) to improve representation and explore potentially useful frequency components, and 3) a pair of convolution layers that leverage large kernels to obtain global range-interact features with stronger shape bias. Compared with the previous frequency methods, the learnable filter in SFE enables FFT to capture both spatial and temporal information, and the energy function in EFE can alleviate the noise accumulated in multiple video frames. Moreover, MFE only contains 0.02M parameters, making it affordable for edge devices.

Besides MFE, we further propose a novel model training method named privilege training (PT) to facilitate the training of VSR models. Motivated by previous research in learning using privilege information (LUPI) [53], we propose to apply a trainable encoding module to encode the privilege information from HR videos and then employ them to obtain a good initialization for VSR models. Sufficient experiments demonstrate the effectiveness of privilege training on both our model and the other VSR models.

With MFE and privilege training, we introduce a novel VSR model referred to as MFPI. Extensive experimental results have demonstrated that MFPI outperforms eighteen previous methods by a clear margin in six VSR benchmarks. For instance, on UDM10, compared with BasicVSR++ [4], which has similar parameters and runtime with MFPI, MFPI achieves 0.36 dB higher PSNR improvements on UDM10 BD degradation, as shown in Figure 1. To sum up, the main contributions in this paper can be summarized as follows:

- We propose multi-frequency representation enhancement (MFE), a module that effectively aggregates information in the frequency domain by operating the spatial- and energy-frequency components.
- We propose privilege training, which encodes the privilege information from the HR videos to boost the performance of VSR models.
- Abundant experiments demonstrate the effectiveness of MFPI on four datasets, including REDS4, Vimeo, Vid4, and UDM10 in both BI and BD degradation.

2. Related Work

2.1. Video Super Resolution

Different from SISR, VSR typically produces higher-quality results by utilizing information from neighboring frames. Recently, learning-based approaches have been highly effective in solving the problem of VSR [16, 4]. The main challenge for VSR is how to correctly fuse auxiliary frames in the presence of dynamic content and object motion. To address this problem, abundant methods have been proposed to explicitly use optical flow [55], deformable convolution [57, 50, 3], and homography [18] to align neighbor frames. However, estimating accurate optical flow or transformation is still challenging in large motions. Recently, some methods have been proposed to learn the spatial correspondence across different frames to tackle this challenge [25, 13]. However, these methods have limited performance in VSR since they cannot capture information between adjacent locations or long-range interactions [47, 41]. To address this problem, we introduce a spatial-frequency representation enhancement branch that models long-range spatial and temporal dependencies.

2.2. Learning in Frequency Domain

Due to the difficulty of distinguishing high-frequency textures from artifacts, fruitful methods [70, 6, 14] have been introduced to decouple them in the frequency space and reconstruct the high-quality textures. For instance, Zhou *et al.* proposed a deep Fourier up-sampling method to recover the resolution in the frequency domain [72]. Guo *et al.* utilized the wavelet technique to simplify the mapping between the LR and HR images, reducing the artificial blocks and recovering the edge‘s information [10]. Motivated by their success, abundant works have tried to apply frequency methods to SR. Dario *et al.* proposed a Fourier space supervision loss to improve restoration results [6]. Qiu *et al.* designed a fine-grained self-attention module in the space-frequency domain, which reconstructs the real visual texture without artifacts [41]. However, the images that are directly transformed into the frequency domain still face obstacles as follows: (1) Decomposing features into different frequency components usually lacks local correlation information in the channel dimension, making it difficult to utilize the useful frequency components [40]. (2) Non-parametric frequency-based methods are easily susceptible to noise and image scale factors [35]. To address the above problems, this paper proposes to combine the advantages of the energy function and frequency-based representations by using DCT with learnable filters.

2.3. Learning Using Privilege Information

Learning using privilege information (LUPI) was first proposed by Vapnik et *et al.* to employ the privilege information provided by a teacher model in support vector machine [53]. Generally speaking, LUPI aims to improve the performance of a model by using privilege information, which is utilized during training but not required during inference. Vapnik *et al.* further utilized LUPI to accelerate the inference speed of SVMs by similarity control and knowledge transfer [52]. Motivated by the success of knowledge distillation [12] on deep neural networks, recently, David *et al.* proposed generalized distillation, which combines knowledge distillation with learning with privilege information theoretically [32]. Recently, label-guided auxiliary was introduced to apply LUPI to point cloud-based 3D detection by using the privilege information from annotations [15]. Lee *et al.* proposed PISR, which introduces LUPI to single image super-resolution by using variational knowledge distillation [24]. Moreover, LUPI has also been utilized in semi-supervised learning [7], domain adaptation [37], speech recognition [36] and so on.

The most related work of the privilege training in our method is PISR [24], which is proposed for knowledge distillation on single image super-resolution. Their main difference is that: (1) PISR is proposed for SISR for model compression while our method is proposed for VSR to improve the quality of restoration. (2) Besides, PISR is a knowledge distillation method which mainly leverages the privilege information with distillation loss. In contrast, our method leverages the privilege information by sharing model weights. (3) PISR implicitly encodes the privilege information via the reconstruction of LR images while our method explicitly encodes privilege information from HR videos with a trainable network. Besides, ablation studies in Table 3 also demonstrate that our method leads to significantly higher performance than PISR.

3. Methodology

3.1. Overall Framework for VSR

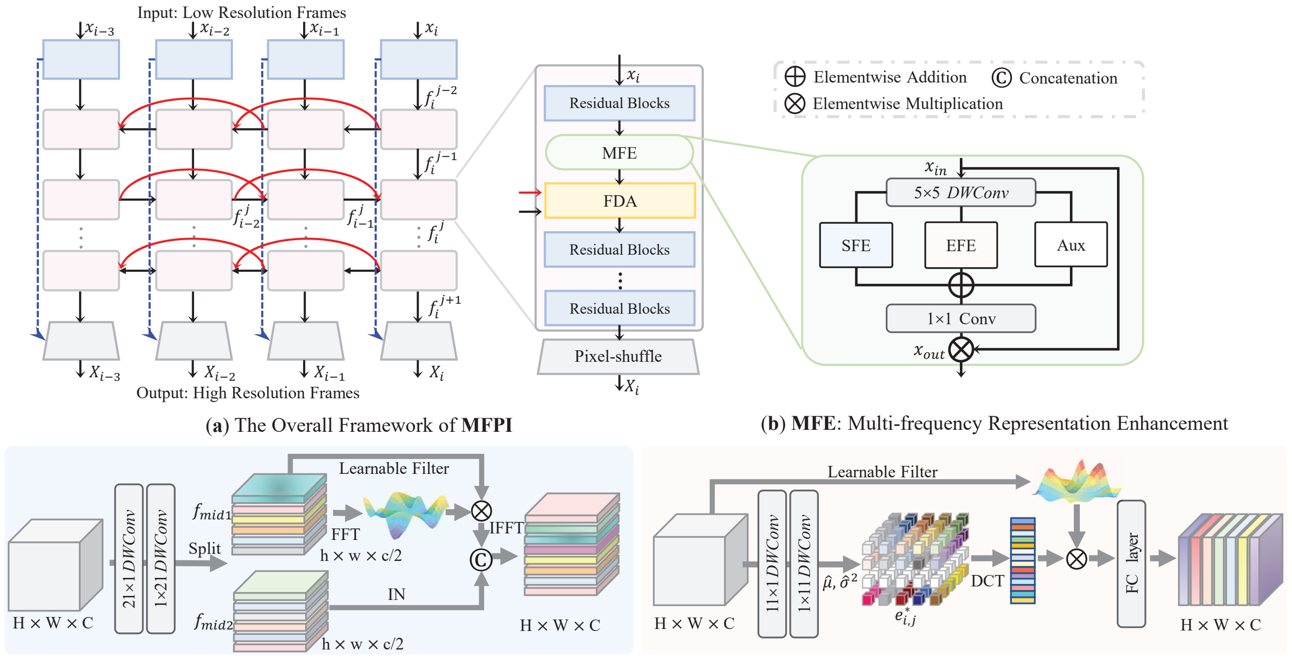
Following the previous work [4], we adopt the recurrent bidirectional propagation as the baseline model. Based on its architecture, we propose a novel VSR model termed MFPI, which consists of a multi-frequency representation enhancement module (MFE) to improve the LR frames representation in the frequency domain, and a novel VSR model training method named privilege training which encodes privilege information from HR frames to facilitate model training. The overall architecture is shown in Figure 2: (1) The VSR model extracts features from the LR frames images to reconstruct the HR frames images. (2) Residual blocks and MFE are employed to extract the feature f_i^j from the LR images, where f_i^j denotes the feature at i_{th} timestamp in the j_{th} propagation branch. (3) The flow-guided deformable alignment block (FDA) aligns the neighboring features f_{i-1}^j and f_{i-2}^j. And the pixel-shuffle operator is employed to perform upsampling $4\times$ scales from LR to HR frames. Besides, learning privilege information boosts the performance of our VSR models during training, as shown in Figure 3.

3.2. MFE: A Multi-Frequency Representation Enhancement Module

As depicted in Figure 2 (b), we design a multi-frequency representation enhancement (MFE) to improve the representation ability and aggregate spatial-temporal information in VSR architecture. MFE is a plug-in-play building block which adopts a multi-branch structure to capture multi-frequency representation [9]. Concretely, it first encodes the input features with a depthwise convolution (*DWConv*) layer [68] and then processes them with three different branches, including 1) a spatial-frequency representation enhancement branch to capture the spatial-frequency information and long-range dependence features, 2) an energy-frequency representation enhancement branch to improve the representation ability and model the inter-channel relationship in the frequency domain, and 3) an auxiliary branch with a pair of 7×1 and 1×7 *DWConv* to inject the inductive bias and facilitate training [8, 31]. Finally, a 1×1 convolution layer aggregates the diverse features from three branches, and the skip connection is employed to facilitate the convergence of MFE [30].

3.2.1 Spatial Frequency Representation Enhancement

As illustrated in Figure 2 (c), the spatial frequency representation enhancement (SFE) branch first employs a pair of *DWConv* to generate the feature map f_{in} from the input $x_{in} \in \mathbb{R}^{H \times W \times C}$. Then, we split f_{in} into two parts $f_{mid1}, f_{mid2} \in \mathbb{R}^{h \times w \times c/2}$ [26, 69]. The FFT layer pro-



(a) The Overall Framework of **MFPI**

(b) **MFE**: Multi-frequency Representation Enhancement

(c) **SFE**: Spatial Frequency Representation Enhancement

(d) **EFE**: Energy Frequency Representation Enhancement

Figure 2: (a) An overview framework of MFPI. The blue dot line denotes the bilinear upsampling operator. The red curve denotes second-order propagation. “FDA“ denotes the flow-guided deformable alignment block [4]. The green block denotes the proposed multi-frequency representation enhancement (MFE). (b) An overview of the MFE. (c) The details of the spatial-frequency representation enhancement branch in MFE extract spatial-level and long-range dependencies. (d) The details of the energy-frequency representation enhancement branch. $e^*_{i,j}$ denotes the energy feature. Note that the ‘FC layer‘ consists of a sequence of the ‘linear-LReLU-linear-LReLU‘ operator. *DWConv* denotes the depthwise convolution [68].

cesses f_{mid1} represents different spatial dependence information as follows:

$$freq = \mathcal{F}\left[f_{mid1}\right] \in \mathbb{C}^{h \times w \times c/2} \tag{1}$$

where c, h, w denote the input channel, height, width of the feature maps, respectively. $\mathcal{F}[\cdot]$ denotes the 2D FFT, $freq$ represents the frequency feature of f_{mid1}. To make SFE adaptive to diverse features [43, 54], we modulate the $freq$ multiplies with a learnable filter $L \in \mathbb{C}^{h \times w \times c/2}$ in the frequency domain: $\tilde{freq} = freq \times L$. The inverse FFT transfers the modulated frequency feature back to the spatial domain:

$$\tilde{f_{mid1}} \leftarrow \mathcal{F}^{-1}\left[\tilde{freq}\right] \tag{2}$$

Finally, the output feature f_{out} is obtained by concatenating with an instance normalization (IN) [51] layer, which prevents covariance shift and preserves the scale information: $f_{out} = Concat\left(\tilde{f_{mid1}}, \text{IN}\left(f_{mid2}\right)\right)$.

It is also worth noting that in our implementation: (1) SFE utilizes a relatively large local window of *DWConv* (*e.g.*, 21 × 1,1 × 21) to capture the long-range spatial dependencies from features. (2) Besides, a learnable Fourier transform with IN can exhibit clear patterns and generalizations in the frequency domain. Hence, the proposed SFE branch tends to be more efficient than directly utilizing Fourier transforms in the spatial-temporal dimension (Table 2b). (3) Considering that FFT has less computation than DFT, we employ FFT in SFE as the frequency transformation to improve model representation. Moreover, the Fourier domain has the nature of global modeling [54]. Hence, computation on frequency bands implies information aggregation over all the spatial locations of the input images. Specifically, in our SFE, FFT is performed over the whole feature map across all the spatial locations, which is equivalent to the global receptive field of the whole feature. (4) SFE splits features into two branches to improve computational efficiency. Specifically, this split operator enables SFE to capture and combine different aspects or representations of the input data. One part focus on the spatial frequency characteristics or patterns, and the other part act as a form of regularization by introducing additional constraints, such as instance normalization. Their outputs can be concatenated to create a richer and more comprehensive representation. This can enhance the SFE's ability to learn complex relationships and improve its overall performance.

Besides, this two-branch design is also similar to the idea of group convolution and CSN [71], which is a well-known technique of efficient neural networks. Please refer to supplementary materials for theoretical analysis.

3.2.2 Energy Frequency Representation Enhancement

Considering that the frequency methods with a learning filter can be more generic and flexible interactions among spatial locations from diverse images, we introduce an energy function to reweight the input feature and utilize the learnable filter with DCT to represent the rich inter-channel relationship [40]. The architecture of the energy frequency representation enhancement (EFE) is illustrated in Figure 2 (d). More specifically, EFE first utilizes a pair of *DWConv* (*e.g.*, 11 × 1,1 × 11) to generate the feature map f. Then, we calculate the mean $\hat{\mu}$ and variance $\hat{\sigma}^2$ of f, and the energy value [62] can be obtained by minimizing the following:

$$e_{i,j} = \frac{4\left(\hat{\sigma}^2+\delta\right)}{\left(t_{i,j}-\hat{\mu}\right)^2+2\hat{\sigma}^2+2\delta} \tag{3}$$

where δ is the hyper-parameter, $e_{i,j}$ denote the energy value of target pixel $t_{i,j}$, $i \in \{0,1,\cdots,H-1\}$, $j \in \{0,1,\cdots,W-1\}$. The energy feature is obtained by grouping all the $e_{i,j}$ across spatial and channel levels, which rescale the energy value to restrict too large energy value:

$$e^*_{i,j} = \text{LeakyReLU}\left(\frac{1}{\sum_{i\neg=0}^{h}\sum_{j\neg=0}^{w} e_{i,j}}\right) \tag{4}$$

To further enrich the representation ability of EFE, we employ the DCT technique to convert the energy features into the frequency domain as follows:

$$\mathcal{F}_c(i,j) = \frac{2}{\sqrt{HW}}\alpha(i)\alpha(j)\sum_{h=0}^{H-1}\sum_{w=0}^{W-1} e^*_{i,j} \tag{5}$$

$$\mathcal{K}^{i,j}_{h,w} = \cos\left(\frac{(2h+1)i\pi}{2H}\right)\cos\left(\frac{(2w+1)j\pi}{2W}\right) \tag{6}$$

$$f_c^{h,w} = \mathcal{K}^{i,j}_{h,w} \times \mathcal{F}_c\left(e^*_{i,j}\right) \tag{7}$$

where $\mathcal{F}_c$ denotes DCT operation, $\mathcal{K}$ denotes the basis function of DCT, $\alpha(x) = 1/\sqrt{2}$ for $x = 0$ and $\alpha(x) = 1$ otherwise [46]. Finally, the output refined feature as follows: $f_{out} = f_c^{h,w} \times x_{i,j}$. Note that an element-wise multiplication between frequency-domain features and the learnable filters improves the representation ability and makes EFE flexible adaptive to different feature inputs. Besides, the FC layer in Figure 2 (d) consists of a sequence of the 'Linear-LeakyReLU-Linear-LeakyReLU' operators, which convert the channel-wise feature into a scalar representation, which reduces the complexity and improves the effectiveness in the frequency domain. Different from the SFE branch, the EFE branch leverages the merits of the energy function and frequency components to improve the representation power and model the relationship among inter-channel dimensions, as shown in Figure 9.

3.3. Privilege Training

Learning using privilege information is a well-known theory in machine learning, which shows that one can improve the performance of machine learning by using kinds of privilege information, which is available in training but not accessible during inference. Motivated by its effectiveness, in this paper, we propose a novel neural network training method for VSR, which is shown in Figure 3. By denoting the LR and and corresponding HR videos as x and y, respectively, the traditional training method can be formulated as $\min \mathbb{E}[|f(x)-y|]$, where $f(\cdot)$ denotes the super-resolution model. In contrast, in our method, we additionally introduce a lightweight neural network $\mathcal{D}$, which aims to encode the privilege information from the HR frames y to a tensor of privilege information $\mathcal{D}(y)$ which has the same shape with the LR frames x. Then, we can denote a mixture of the LR images and the privilege information as $x^* = x + \gamma \cdot \mathcal{D}(y)$, where $\gamma \in [0,1]$ is a parameter to balance the two items. During the training period, we gradually decrease γ from 1 to 0. Hence, the VSR model is able to firstly learn a good representation by using the privilege information, and then gradually get rid of it. Note that since γ is decreased to 0, the VSR model does not require any privilege information during inference. In summary, its objective of privilege training can be written as

$$\arg\min_{f,\mathcal{D}} \mathbb{E}[|f(x^*)-y| + |\mathcal{D}(y)|] \tag{8}$$

where the first loss item is the common reconstruction loss for VSR training. The second loss item is a regularization item which reduces the energy of privilege information to prevent the model from overusing the privilege information. Note that the regularization item is indispensable since without this item the VSR model may degenerate into a naive auto-encoder which reconstructs the HR images totally based on the privilege information. Note that $\mathcal{D}$ in our implementation is a stack of three convolutional layers followed with LeakyReLU.

Why does privilege training work? The effectiveness of privilege information can be understood from the perspective of weight initialization. During the early training iterations, the model can quickly converge to a flat minima by using the privilege information. Then, by reducing γ, the privilege information gradually makes a less influence to the VSR model, and hence the VSR model can converge

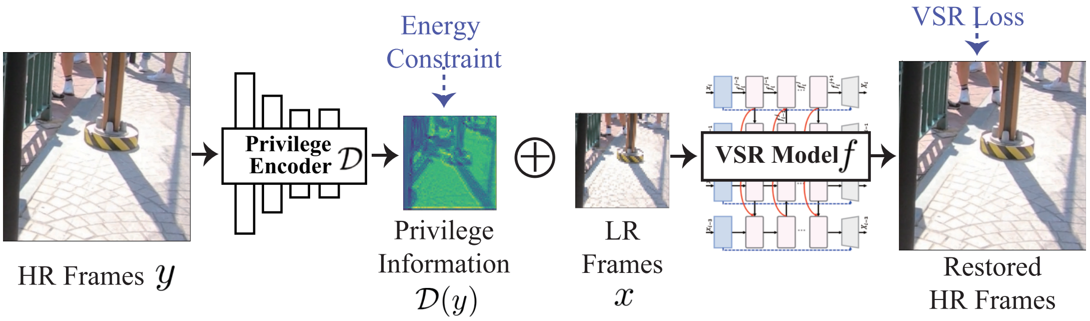


Figure 3: The paradigm of the proposed learning using privilege information on VSR. ⊕ denotes a weighted addition between the privilege information and the LR frames.

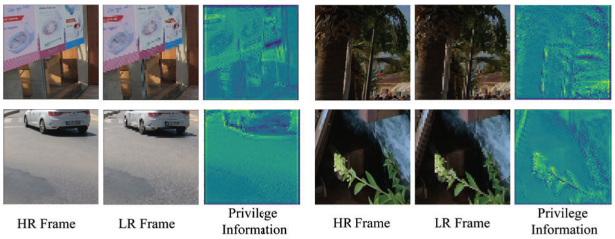


Figure 4: Visualization of the privilege information.

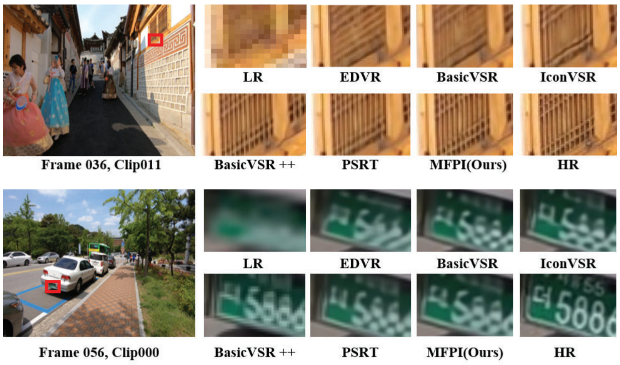


Figure 5: Challenging scenario on REDS4 [38]. MFPI generates sharp edge in repeated structures of the windows and sharper texts.

stably. Secondly, the success of our method can also be understood as a special case of knowledge distillation, where the teacher model is trained with privilege information, the student model is trained with only the LR frames, and the knowledge distillation is applied by sharing the weights of the student and the teacher. Besides, Figure 4 visualizes the privilege information encoded during the training period. It is observed that most of privilege information concentrates on the pixels belonging to the objects and their edges, which are more important in the assessment of image quality. These privilege information is utilized to help VSR models to converge at the early training period.

4. Experiment

4.1. Experimental Setup

Our model is trained for each task with 6×10^5 iterations using randomly cropped patches. We apply data augmentation techniques such as random horizontal and vertical flipping and 90° rotation. The Adam optimizer [22] is used with an initial learning rate of 2×10^{-4}, which is steadily decreased to 1×10^{-7} with cosine annealing learning rate decay [33]. δ is set to 1×10^{-6} [62]. We employ SPyNet to predict optical flow in videos [42]. Each branch contains 7 residual blocks and 64 feature channels. The batch size is set to 4, and the input low-resolution (LR) frames are of patch size 64×64.

Peak Signal-to-Noise Ratio (PSNR) and Structural Similarity Index Measure (SSIM) [58] between ground truth and restored videos are utilized as the metrics in our quantitative comparisons. We follow [4] to utilize REDS [38], Vimeo-90K [60] as training set, and REDS4, Vimeo-90K-T [60], Vid4 [28], UDM10 [65] , and Vimeo-90K as the test set. REDSval4 is utilized as the validation set. All models are tested with $4\times$ downsampling via two degradations – Bicubic (BI) and Blur Downsampling (BD). All experiments are implemented with PyTorch [39] and on two NVIDIA Tesla A100 GPUs.

4.2. Comparisons with State-of-the-Art Methods

Qualitative Comparison We show the comparison of visual results in Figure. 5, 6, 7 and 8. It is observed that the results of prior works usually suffer from incomplete artifacts or blurred regions in the image. In contrast, our model restores sharper edges and finer detailed textures of the number plate in Figure 5 and the display board in Figure 6. Besides, MFPI is able to recover the edge of the small object in Figure 6. As shown in Figure 7 and 8, MFPI also shows

Table 1: Quantitative comparison (PSNR/SSIM). All results are calculated on Y-channel except REDS4 [38] (RGB-channel). The runtime is computed on an LR size of 180×320. A 4× upsampling is performed following previous studies. Blanked entries correspond to results not reported in previous works. Numbers in bold indicate the best performance.

Model	Params	Runtime	BI degradation			BD degradation		
	(M)	(ms)	REDS4 [38]	Vimeo-90K-T [60]	Vid4[28]	UDM10 [65]	Vimeo-90-T [60]	Vid4 [28]
Bicubic	-	-	26.14/0.7292	31.32/0.8684	23.78/0.6347	28.47/0.8253	31.30/0.8687	21.80/0.5246
VESPCN [1]	-	-	-	-	25.35/0.7557	-	-	-
SPMC [49]	-	-	-	-	25.88/0.7752	-	-	-
TOFlow [61]	-	-	27.98/0.7990	33.08/0.9054	25.89/0.7651	36.26/0.9438	34.62/0.9212	-
FRVSR [45]	5.1	137	-	-	-	37.09/0.9522	35.64/0.9319	26.69/0.8103
DUF [21]	5.8	974	28.63/0.8251	-	-	38.48/0.9605	36.87/0.9447	27.38/0.8329
RBPN [11]	12.2	1507	30.09/0.8590	37.07/0.9435	27.12/0.8180	38.66/0.9596	37.20/0.9458	-
EDVR-M [57]	3.3	118	30.53/0.8699	37.09/0.9446	27.10/0.8186	39.40/0.9663	37.33/0.9484	27.45/0.8406
EDVR [57]	20.6	378	31.09/0.8800	37.61/0.9489	27.35/0.8264	39.89/0.9686	37.81/0.9523	27.85/0.8503
PFNL [65]	3.0	295	29.63/0.8502	36.14/0.9363	26.73/0.8029	38.74/0.9627	-	27.16/0.8355
MuCAN [27]	-	-	30.88/0.8750	37.32/0.9465	-	-	-	-
TGA [18]	5.8	384	-	-	-	-	37.59/0.9516	27.63/0.8423
RLSP [5]	4.2	49	-	-	-	38.48/0.9606	36.49/0.9403	27.48/0.8388
RSDN [17]	6.2	94	-	-	-	39.35/0.9653	37.23/0.9471	27.92/0.8505
RRN [19]	3.4	45	-	-		38.96/0.9644	-	27.69/0.8488
BasicVSR [2]	6.3	63	31.42/0.8909	37.18/0.9450	27.24/0.8251	39.96/0.9694	37.53/0.9498	27.96/0.8553
IconVSR [2]	8.7	70	31.67/0.8948	37.47/0.9476	27.39/0.8279	40.03/0.9694	37.84/0.9524	28.04/0.8570
BasicVSR++ [4]	7.3	77	32.39/0.9069	37.79/0.9530	27.79/0.8400	40.72/0.9722	38.21/0.9550	29.04/0.8753
PSRT [47]	13.4	812	32.72/0.9106	38.27/**0.9536**	28.07/**0.8485**	-	-	-
MFPI (Ours)	7.3	76	**32.81/0.9106**	**38.28**/0.9534	**28.11**/0.8481	**41.08/0.9741**	**38.70/0.9579**	**29.34/0.8781**

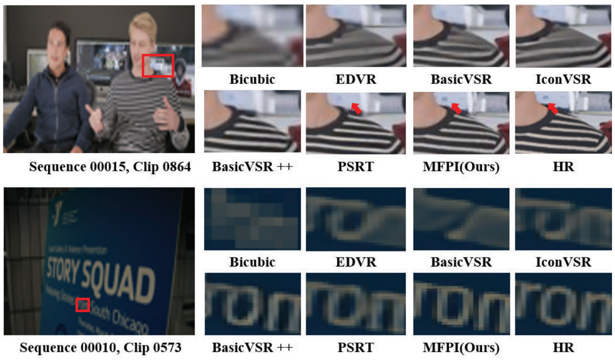


Figure 6: Challenging scenario on Vimeo-90K-T [60]. Compared to the other methods, Our method more effectively preserves while recovering the strip-like features in the skirts.

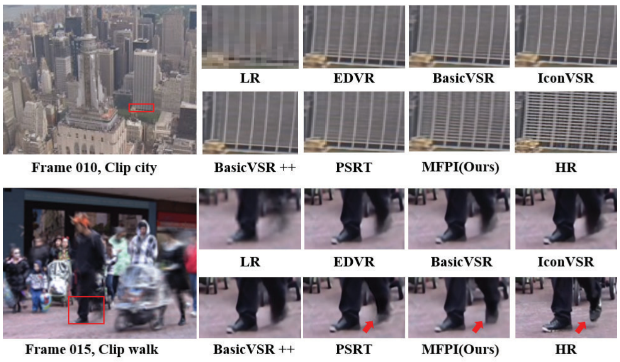


Figure 7: Challenging scenario on Vid4 [28]. MFPI recovers the clear textures and reduces the produce blurry results.

significantly better performance than prior works when the camera and the foreground objects are moving, which is a more challenging case in VSR. Moreover, MFPI can recover the scene details more faithfully than other methods, such as pedestrians' feet and text in Figure 7. More visual comparisons are shown in the supplementary material.

Quantitative Comparison The quantitative results on six test sets with scaling factors 4 are shown in Table 1. It is observed that: (1) Compared with existing approaches that only perform well on a certain dataset, our MFPI achieves the best performance in all the datasets for both BI and BD degradation. (2) Compared with BasicVSR++ which has very similar parameters and speed to our method, 0.42 dB, 0.49 dB and 0.32 dB PSNR improvements can be achieved on the three datasets in BI degradation experiments, respectively. (3) Compared with PSRT-recurrent which has slightly lower PSNR than our method, our model only has around 54% parameters, which indicates better efficiency. These observations indicate that our method significantly outperforms the previous VSR models.

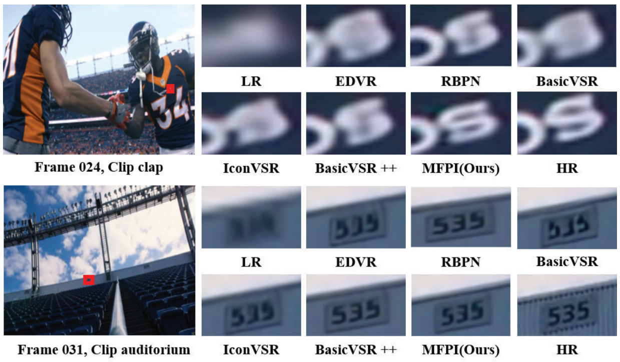


Figure 8: Challenging scenario on UDM10 [65]. Our method produces much clearer text and sharper edges.

4.3. Ablation Studies

This subsection gives a series of ablation studies to study the effectiveness of each component of MFPI. The evaluations were performed on the REDS4 dataset [38] trained on image patches of size 180 × 320 for 6 × 10^5 iterations. BasicVSR++ is utilized as the baseline.

A. Individual components. An ablation study is conducted by progressively adding SFE, EFE and the auxiliary branch to study the influence of different network blocks in our MFPI. Besides, we also investigate the effectiveness of the proposed privilege training and compare it with knowledge distillation (KD), which is another popular neural network training method. As shown in Table 2a: (a) 0.16 dB and 0.13 dB PSNR improvements can be obtained by using SFE and EFE individually. Besides, 0.02 dB further PSNR improvements can be obtained by combining the two methods with the auxiliary branch. (b) Privilege training leads to 0.12 PSNR improvements while KD leads to 0.38 dB PSNR reduction, indicating the traditional model training method fails in VSR while our method leads to significant performance benefits. As Table 2a suggests, combining all the components yields the best performance vs. complexity tradeoff (32.81 dB, 7.34 M, and 281.25 G) compared to employing single or partial components.

B. Ablation on SFE. We further study the performance of our proposed SFE in Table 2b with different variants: only FFT, FFT with learnable filter, the split input with FL and original feature, the split operator with FL and batch normalization, the split operator with FL and instance normalization. Then, we incrementally added different *DWConv* layers on top of the existing components, such as 3 × 3, 7 × 7, 11 × 11, and 21 × 21. Besides, we only substituted the FFT with the original feature while keeping the LF, IN, and *DWConv* 21×21 (-w/o FFT). Experimental results show that: (i) Only adding FFT to Base leads to 0.04 PSNR drop, indicating that only utilizing Fourier transform cannot boost performance. (ii) Using the learnable filter leads to 0.07 dB improvements, which account for 41% performance benefits, indicating the learnable filter can improve the representation ability of FFT. (iii) Splitting the features into FFT and IN branches leads to 0.07 dB PSNR benefits, indicating that IN can help the model preserve scale information and prevent covariance shifts. (iv) *DWConv* with 21×21 kernel performs better than other kernels, indicating a larger kernel is necessary. (v) SFE without FFT achieves only 31.93 dB PSNR, indicating FFT plays an essential role in SFE.

C. Ablation on EFE. To evaluate different DCT basis functions in the EFE, we have compared our method with four different DCT variants, which are DCT initialization with fixed coefficients (DF), DCT initialization with learnable filter (DL), DF and DL added energy function, respectively. Experimental results show that: (i) Our method achieves the best PSRN, which demonstrates the effectiveness of using a learnable filter and the energy function. (ii) Removing DCT from EFE leads to 0.43 dB PSNR drop, which indicates EFE is necessary. (iii) EFE with *DWConv* 11x11 kernel achieves the best PSNR than other *DWConv* kernels, indicating that a relatively larger kernel tends to achieve better performance.

D. Privilege Training As shown in Table 2a, 0.12 dB PSNR improvements can be obtained by applying PT to our model, which is 0.50 higher than KD. To further study its effectiveness, we have compared PT with the other six popular model training methods in Table 3 on BasicVSR [2]. It is observed that four of the previous methods do not bring any performance benefits but lead to significant PSNR and SSIM drop. PISR leads to 0.04 dB PSNR and 0.0009 SSIM improvements, which are 0.14 dB and 0.0022 lower than PT in PSNR and SSIM, respectively. Besides, PT can be easily implemented without a pre-trained teacher model.

E. Feature Visualization Visualization of intermediate features in MFPI is shown in Figure 9. It is observed that: (1) The feature map significantly reduces noise after MFE, indicating that MFE effectively suppresses noise information. (2) SFE can capture object motion and extract spatial features in the frequency domain. (3) EFE focuses on the fine-grained texture and details of the input feature, as shown in the red arrow and circle. (4) Moreover, the MFE effectively fuses the multi-frequency feature and enhances the representation ability.

5. Conclusion

In this paper, we have proposed a novel VSR model referred to as MFPI, which consists of a multi-frequency representation enhancement module (MFE) and a privilege training method. MFE is utilized to aggregate informa-

Table 2: Ablation studies. Subtables 2a, 2b, and 2c contain components that are defined in Sec. 3.2 and 3.3. We conduct ablation experiments on various components of SFE and EFE, such as FFT/DCT, split operator, learnable filter, BN/IN, energy function, and different coefficients. We also evaluate the effects of different kernel sizes using *DWConv*. Additionally, we employ knowledge distillation [12] or privilege training (PT) to improve our results.

Base	MFE			Training		PSNR	Params	FLOPs
	SFE	EFE	Aux	KD	PT	(dB)	(M)	(G)
✓						32.39	7.32	280.59
✓	✓					32.55	7.34	280.76
✓		✓				32.52	7.33	280.68
✓	✓	✓				32.67	7.34	280.86
✓	✓	✓	✓			32.69	7.34	281.25
✓	✓	✓	✓	✓		32.31	7.34	281.25
✓	✓	✓	✓		✓	**32.81**	7.34	281.25

(a) Individual components.

Variant	PSNR
Base + FFT	32.35
Base + FFT w/. Learnable filter	32.48
Base + FFT w/. Learnable filter + original feature	32.49
Base + FFT w/. Learnable filter + BN	32.43
Base + FFT w/. Learnable filter + IN	32.52
Base + FFT w/. Learnable filter + IN + *DWConv* 3 × 3	32.48
Base + FFT w/. Learnable filter + IN + *DWConv* 7 × 7	32.52
Base + FFT w/. Learnable filter + IN + *DWConv* 11 × 11	32.51
Base + FFT w/. Learnable filter + IN + *DWConv* 21 × 21 (Our SFE)	**32.55**
Base + w/. Learnable filter + IN + *DWConv* 21 × 21	31.93

(b) Effects of the SFE branch.

Variant	PSNR
Base + Energy function	32.40
Base + DCT w/. Fixed coefficients	32.32
Base + DCT w/. Learnable filter	32.41
Base + DCT w/. Fixed coefficients + Energy function	32.37
Base + DCT w/. Learnable filter + Energy function	32.50
Base + DCT w/. Learnable filter + Energy function + *DWConv* 3 × 3	32.42
Base + DCT w/. Learnable filter + Energy function + *DWConv* 7 × 7	32.40
Base + DCT w/. Learnable filter + Energy function + *DWConv* 11 × 11 (Our EFE)	**32.52**
Base + DCT w/. Learnable filter + Energy function + *DWConv* 21 × 21	32.43
Base + w/. Learnable filter + Energy function + *DWConv* 11 × 11	32.07

(c) Effects of the EFE branch.

Table 3: Comparison between the proposed privilege training and previous training methods with BasicVSR on [38].

Method	Needs Teacher?	PSNR	SSIM
BasicVSR	✓	31.58	0.8940
+ Deep Supervision [23]	×	31.58	0.8941
+ Self-KD [67]	×	31.47	0.8914
+ Hinton KD [12]	✓	31.26	0.8883
+ FitNet [44]	✓	31.52	0.8934
+ PISR [24]	✓	31.62	0.8949
+ Ours	×	**31.76**	**0.8971**

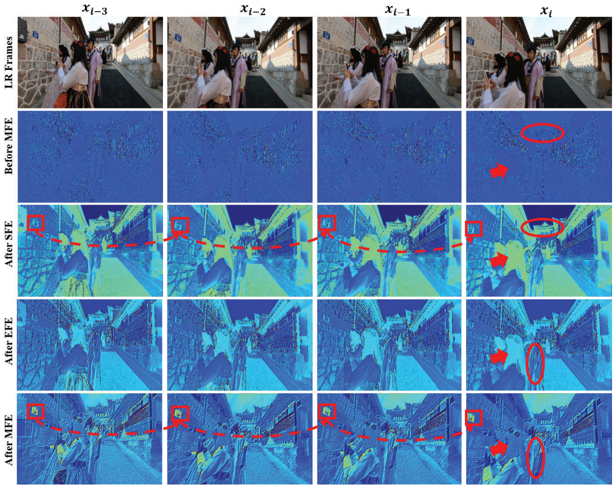


Figure 9: Visualization of the features in MFPI on images from REDS4[38]. The red dot curve denotes the change in spatial information at different locations of the same object.

tion in the frequency domain by operating the spatial- and energy-frequency components via SFE and EFE. It enables CNN-based VSR models to capture long-range dependencies with very minor additional parameters and computations. Besides, we also propose to improve the performance of VSR from a training perspective by introducing privilege training, which contributes orthogonal effects to our frequency modules. We hope that this paper may promote more research that studies VSR from the perspective of frequency methods and model training methods.

6. Future Work

In the future, exploring MFPI in a real-world scenario would be an exciting direction, which offers a new valuable perspective on our proposed MFEI performance. We speculate that our frequency methods will be useful for real-scenario tasks and a broader investigation of it has provided additional depth to our study. We will continue to (1) evaluate MFPI on more challenging datasets such as RealVSR [63], and (2) deploy MFPI on edge devices and evaluate its performance. We are actively working on conducting the requested experiments that can further strengthen our conclusions.

Acknowledgement

The authors acknowledge the financial support from the National Key R&D Program of China $(No.2021ZD0113805)$. We appreciate the comments from the seminar participants at the Center for Deep Learning of Computer Vision Research at China Agricultural University, which improves the manuscript significantly.

References

[1] Jose Caballero, Christian Ledig, Andrew Aitken, Alejandro Acosta, Johannes Totz, Zehan Wang, and Wenzhe Shi. Real-time video super-resolution with spatio-temporal networks and motion compensation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4778–4787, 2017. 7

[2] 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 (CVPR)*, pages 4947–4956, 2021. 1, 7, 8

[3] Kelvin CK Chan, Xintao Wang, Ke Yu, Chao Dong, and Chen Change Loy. Understanding deformable alignment in video super-resolution. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 973–981, 2021. 1, 2

[4] 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 (CVPR)*, pages 5972–5981, 2022. 2, 3, 4, 6, 7

[5] Dario Fuoli, Shuhang Gu, and Radu Timofte. Efficient video super-resolution through recurrent latent space propagation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*, pages 3476–3485. IEEE, 2019. 7

[6] Dario Fuoli, Luc Van Gool, and Radu Timofte. Fourier space losses for efficient perceptual image super-resolution. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 2360–2369, 2021. 2, 3

[7] Chen Gong, Xiaojun Chang, Meng Fang, and Jian Yang. Teaching semi-supervised classifier via generalized distillation. In *IJCAI*, pages 2156–2162, 2018. 3

[8] Jiaqi Gu, Hyoukjun Kwon, Dilin Wang, Wei Ye, Meng Li, Yu-Hsin Chen, Liangzhen Lai, Vikas Chandra, and David Z Pan. Multi-scale high-resolution vision transformer for semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12094–12103, 2022. 3

[9] Meng-Hao Guo, Cheng-Ze Lu, Qibin Hou, Zhengning Liu, Ming-Ming Cheng, and Shi-Min Hu. Segnext: Rethinking convolutional attention design for semantic segmentation. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022. 3

[10] Tiantong Guo, Hojjat Seyed Mousavi, Tiep Huu Vu, and Vishal Monga. Deep wavelet prediction for image super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 104–113, 2017. 2

[11] Muhammad Haris, Gregory Shakhnarovich, and Norimichi Ukita. Recurrent back-projection network for video super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3897–3906, 2019. 1, 7

[12] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2014. 3, 9

[13] Mengshun Hu, Kui Jiang, Liang Liao, Jing Xiao, Junjun Jiang, and Zheng Wang. Spatial-temporal space hand-in-hand: Spatial-temporal video super-resolution via cycle-projected mutual learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3574–3583, 2022. 2

[14] Huaibo Huang, Ran He, Zhenan Sun, and Tieniu Tan. Wavelet-srnet: A wavelet-based cnn for multi-scale face super resolution. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 1689–1697, 2017. 2

[15] Yaomin Huang, Xinmei Liu, Yichen Zhu, Zhiyuan Xu, Chaomin Shen, Zhengping Che, Guixu Zhang, Yaxin Peng, Feifei Feng, and Jian Tang. Label-guided auxiliary training improves 3d object detector. *arXiv preprint arXiv:2207.11753*, 2022. 3

[16] Yan Huang, Wei Wang, and Liang Wang. Video super-resolution via bidirectional recurrent convolutional networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(4):1015–1028, 2017. 2

[17] Takashi Isobe, Xu Jia, Shuhang Gu, Songjiang Li, Shengjin Wang, and Qi Tian. Video super-resolution with recurrent structure-detail network. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 645–660. Springer, 2020. 7

[18] Takashi Isobe, Songjiang Li, Xu Jia, Shanxin Yuan, Gregory Slabaugh, Chunjing Xu, Ya-Li Li, Shengjin Wang, and Qi Tian. Video super-resolution with temporal group attention. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8008–8017, 2020. 1, 2, 7

[19] Takashi Isobe, Fang Zhu, Xu Jia, and Shengjin Wang. Revisiting temporal modeling for video super-resolution. *arXiv preprint arXiv:2008.05765*, 2020. 7

[20] Kui Jiang, Zhongyuan Wang, Peng Yi, and Junjun Jiang. Hierarchical dense recursive network for image super-resolution. *Pattern Recognition*, 107:107475, 2020. 1

[21] Younghyun Jo, Seoung Wug Oh, Jaeyeon Kang, and Seon Joo Kim. Deep video super-resolution network using dynamic upsampling filters without explicit motion compensation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3224–3232, 2018. 7

[22] Diederik P Kingma and Jimmy Lei Ba. Adam: A method for stochastic gradient descent. In *Proceedings of the International Conference on Learning Representations (ICLR)*, pages 1–15, 2015. 6

[23] Chen-Yu Lee, Saining Xie, Patrick Gallagher, Zhengyou Zhang, and Zhuowen Tu. Deeply-supervised nets. In *Artificial Intelligence and Statistics*, pages 562–570, 2015. 9

[24] Wonkyung Lee, Junghyup Lee, Dohyung Kim, and Bumsub Ham. Learning with privileged information for efficient image super-resolution. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 465–482. Springer, 2020. 3, 9

[25] Dongxu Li, Chenchen Xu, Kaihao Zhang, Xin Yu, Yiran Zhong, Wenqi Ren, Hanna Suominen, and Hongdong Li. Arvo: Learning all-range volumetric correspondence for video deblurring. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7721–7731, 2021. 2

[26] Fei Li, Lingfeng Shen, Yang Mi, and Zhenbo Li. Drcnet: Dynamic image restoration contrastive network. In *European Conference on Computer Vision (ECCV)*, pages 514–532. Springer, 2022. 3

[27] Wenbo Li, Xin Tao, Taian Guo, Lu Qi, Jiangbo Lu, and Jiaya Jia. Mucan: Multi-correspondence aggregation network for video super-resolution. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 335–351. Springer, 2020. 7

[28] Ce Liu and Deqing Sun. On bayesian adaptive video super resolution. *IEEE transactions on pattern analysis and machine intelligence*, 36(2):346–360, 2013. 6, 7

[29] 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 (CVPR)*, pages 5687–5696, 2022. 2

[30] Fenglin Liu, Xuancheng Ren, Zhiyuan Zhang, Xu Sun, and Yuexian Zou. Rethinking skip connection with layer normalization. In *Proceedings of the 28th International Conference on Computational Linguistics (COLING)*, pages 3586–3598, 2020. 3

[31] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11976–11986, 2022. 3

[32] David Lopez-Paz, Léon Bottou, Bernhard Schölkopf, and Vladimir Vapnik. Unifying distillation and privileged information. *Proceedings of the International Conference on Learning Representations (ICLR)*, 2016. 3

[33] I Loshchilov and F Hutter. Sgdr: Stochastic gradient descent with warm restarts. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2017. 6

[34] Jianping Luo, Shaofei Huang, and Yuan Yuan. Video super-resolution using multi-scale pyramid 3d convolutional networks. In *Proceedings of the ACM International Conference on Multimedia*, pages 1882–1890, 2020. 1

[35] Morteza Mardani, Guilin Liu, Aysegul Dundar, Shiqiu Liu, Andrew Tao, and Bryan Catanzaro. Neural ffts for universal texture image synthesis. *Advances in Neural Information Processing Systems (NeurIPS)*, 33:14081–14092, 2020. 3

[36] Konstantin Markov and Tomoko Matsui. Robust speech recognition using generalized distillation framework. In *Interspeech*, pages 2364–2368, 2016. 3

[37] Saeid Motiian and Gianfranco Doretto. Information bottleneck domain adaptation with privileged information for visual recognition. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 630–647. Springer, 2016. 3

[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 (CVPRW)*, 2019. 6, 7, 8, 9

[39] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 8024–8035. Curran Associates, Inc., 2019. 6

[40] Zequn Qin, Pengyi Zhang, Fei Wu, and Xi Li. Fcanet: Frequency channel attention networks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 783–792, 2021. 2, 3, 5

[41] Zhongwei Qiu, Huan Yang, Jianlong Fu, and Dongmei Fu. Learning spatiotemporal frequency-transformer for compressed video super-resolution. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2022. 2, 3

[42] Anurag Ranjan and Michael J Black. Optical flow estimation using a spatial pyramid network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4161–4170, 2017. 6

[43] Yongming Rao, Wenliang Zhao, Zheng Zhu, Jiwen Lu, and Jie Zhou. Global filter networks for image classification. *Advances in Neural Information Processing Systems (NeurIPS)*, 34:980–993, 2021. 4

[44] Adriana Romero, Nicolas Ballas, Samira Ebrahimi Kahou, Antoine Chassang, Carlo Gatta, and Yoshua Bengio. Fitnets: Hints for thin deep nets. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2015. 9

[45] Mehdi SM Sajjadi, Raviteja Vemulapalli, and Matthew Brown. Frame-recurrent video super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6626–6634, 2018. 1, 7

[46] Xing Shen, Jirui Yang, Chunbo Wei, Bing Deng, Jianqiang Huang, Xian-Sheng Hua, Xiaoliang Cheng, and Kewei Liang. Dct-mask: Discrete cosine transform mask representation for instance segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8720–8729, 2021. 5

[47] Shuwei Shi, Jinjin Gu, Liangbin Xie, Xintao Wang, Yujiu Yang, and Chao Dong. Rethinking alignment in video super-resolution transformers. *arXiv preprint arXiv:2207.08494*, 2022. 2, 7

[48] Huihui Song, Yutong Jin, Yong Cheng, Bo Liu, Dong Liu, and Qingshan Liu. Learning interlaced sparse sinkhorn matching network for video super-resolution. *Pattern Recognition*, 124:108475, 2022. 1

[49] Xin Tao, Hongyun Gao, Renjie Liao, Jue Wang, and Jiaya Jia. Detail-revealing deep video super-resolution. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 4472–4480, 2017. 7

[50] 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 (CVPR)*, pages 3357–3366, 2020. 1, 2

[51] Dmitry Ulyanov, Andrea Vedaldi, and Victor Lempitsky. Instance normalization: The missing ingredient for fast stylization. *arXiv preprint arXiv:1607.08022*, 2016. 4

[52] Vladimir Vapnik, Rauf Izmailov, et al. Learning using privileged information: similarity control and knowledge transfer. *J. Mach. Learn. Res.*, 16(1):2023–2049, 2015. 3

[53] Vladimir Vapnik and Akshay Vashist. A new learning paradigm: Learning using privileged information. *Neural networks*, 22(5-6):544–557, 2009. 2, 3

[54] Chenyang Wang, Junjun Jiang, Zhiwei Zhong, and Xianming Liu. Spatial-frequency mutual learning for face super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 22356–22366, 2023. 4

[55] Longguang Wang, Yulan Guo, Li Liu, Zaiping Lin, Xinpu Deng, and Wei An. Deep video super-resolution using hr optical flow estimation. *IEEE Transactions on Image Processing*, 29:4323–4336, 2020. 2

[56] Peihao Wang, Wenqing Zheng, Tianlong Chen, and Zhangyang Wang. Anti-oversmoothing in deep vision transformers via the fourier domain analysis: From theory to practice. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2022. 2

[57] 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 (CVPRW)*, 2019. 2, 7

[58] 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. 6

[59] Zhongyuan Wang, Peng Yi, Kui Jiang, Junjun Jiang, Zhen Han, Tao Lu, and Jiayi Ma. Multi-memory convolutional neural network for video super-resolution. *IEEE Transactions on Image Processing*, 28(5):2530–2544, 2018. 1

[60] Tianfan Xue, Baian Chen, Jiajun Wu, Donglai Wei, and William T Freeman. Video enhancement with task-oriented flow. *International Journal of Computer Vision*, 127(8):1106–1125, 2019. 6, 7

[61] Tianfan Xue, Baian Chen, Jiajun Wu, Donglai Wei, and William T Freeman. Video enhancement with task-oriented flow. *International Journal of Computer Vision*, 127(8):1106–1125, 2019. 7

[62] Lingxiao Yang, Ru-Yuan Zhang, Lida Li, and Xiaohua Xie. Simam: A simple, parameter-free attention module for convolutional neural networks. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 11863–11874, 2021. 5, 6

[63] Xi Yang, Wangmeng Xiang, Hui Zeng, and Lei Zhang. Real-world video super-resolution: A benchmark dataset and a decomposition based learning scheme. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 4781–4790, 2021. 1, 9

[64] Ting Yao, Yingwei Pan, Yehao Li, Chong-Wah Ngo, and Tao Mei. Wave-vit: Unifying wavelet and transformers for visual representation learning. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 328–345, 2022. 2

[65] 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 (ICCV)*, pages 3106–3115, 2019. 1, 6, 7, 8

[66] Yingchen Yu, Fangneng Zhan, Shijian Lu, Jianxiong Pan, Feiying Ma, Xuansong Xie, and Chunyan Miao. Wavefill: A wavelet-based generation network for image inpainting. In *Proceedings of the IEEE/CVF international conference on computer vision (ICCV)*, pages 14114–14123, 2021. 2

[67] Linfeng Zhang, Jiebo Song, Anni Gao, Jingwei Chen, Chenglong Bao, and Kaisheng Ma. Be your own teacher: Improve the performance of convolutional neural networks via self distillation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (CVPR)*, pages 3713–3722, 2019. 9

[68] Pengfei Zhang, Eric Lo, and Baotong Lu. High performance depthwise and pointwise convolutions on mobile devices. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 6795–6802, 2020. 3, 4

[69] Xiaole Zhao, Yulun Zhang, Tao Zhang, and Xueming Zou. Channel splitting network for single mr image super-resolution. *IEEE transactions on image processing*, 28(11):5649–5662, 2019. 3

[70] Zixiang Zhao, Jiangshe Zhang, Shuang Xu, Zudi Lin, and Hanspeter Pfister. Discrete cosine transform network for guided depth map super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5697–5707, 2022. 2

[71] Zhisheng Zhong, Tiancheng Shen, Yibo Yang, Zhouchen Lin, and Chao Zhang. Joint sub-bands learning with clique structures for wavelet domain super-resolution. *Advances in Neural Information Processing Systems (NeurIPS)*, 31, 2018. 2, 5

[72] Man Zhou, Yu Hu, Jie Huang, Feng Zhao, Jinwei Gu, Chen Change Loy, Deyu Meng, and Chongyi Li. Deep fourier up-sampling. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022. 2

[73] Xiaobin Zhu, Zhuangzi Li, Xiao-Yu Zhang, Changsheng Li, Yaqi Liu, and Ziyu Xue. Residual invertible spatio-temporal network for video super-resolution. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 5981–5988, 2019. 1