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 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].

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
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 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 $f^t_i$ from the LR images, where $f^t_i$ 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 $f^t_{i-1}$ and $f^t_{i-2}$. 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 mult-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 $7 \times 1$ and $1 \times 7$ DWConv to inject the inductive bias and facilitate training [8, 31]. Finally, a $1 \times 1$ 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 $f_i$ from the input $x_i \in \mathbb{R}^{H \times W \times C}$. Then, we split $f_i$ into two parts $f_{mid1}, f_{mid2} \in \mathbb{R}^{h \times w \times c/2}$ [26, 69]. The FFT layer pro-
fers the modulated frequency feature back to the spatial frequency domain: $\tilde{DWConv}$

SFE utilizes a relatively large local window of $\text{DWConv}$ (e.g., $21 \times 1,1 \times 21$) 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 focuses on the spatial frequency characteristics or patterns, and the other part acts 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.

$\text{EFE}$: Energy Frequency Representation Enhancement

$\text{FDA}$: Flow-guided deformable alignment block

5x5 $\text{DWConv}$

0 $\text{FC}$ layer

0 $\text{Aux}$

$\text{H} \times \text{W} \times \text{C}$

$\text{Input: Low Resolution Frames}$

$\text{Output: High Resolution Frames}$

$\text{(a) The Overall Framework of MFPI}$

$\text{(b) MFE: Multi-frequency Representation Enhancement}$

$\text{(c) SFE: Spatial Frequency Representation Enhancement}$

$\text{(d) EFE: Energy Frequency Representation Enhancement}$

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. $e_{i,j}^f$ denotes the energy feature. Note that the ‘FC layer’ consists of a sequence of the ‘linear-LReLU-linear-LReLU’ operator. $\text{DWConv}$ denotes the depthwise convolution [68].
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 \(\text{DWConv} \) (e.g., \(11 \times 1,1 \times 1\)) to generate the feature map \(f\). Then, we calculate the mean \(\mu\) and variance \(\hat{\sigma}^2\) of \(f\), and the energy value \([62]\) can be obtained by minimizing the following:

\[
e_{i,j} = \frac{4 (\hat{\sigma}^2 + \delta)}{(\hat{i}_{i,j} - \mu)^2 + 2\hat{\sigma}^2 + 2\delta} \tag{3}
\]

where \(\delta\) 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:

\[
e^*_{i,j} = \text{LeakyReLU}\left(\frac{\alpha}{\sum_{i=0}^{H} \sum_{j=0}^{W} e_{i,j}}\right) \tag{4}
\]

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

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

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

\[
f^c_{h,w} = f^h_{c,w} \times \mathcal{F}_c(e^*_{i,j}) \tag{7}
\]

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: \(f_{\text{out}} = f^H_c \times x_{i,j}\). 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-Linerar-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 \(f(.)\) 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 \(\gamma\) 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 \(\gamma\) 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 oversusing 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 \(\gamma\), 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.

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 $6 \times 10^5$ 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 $2 \times 10^{-4}$, which is steadily decreased to $1 \times 10^{-7}$ with cosine annealing learning rate decay [33]. $\delta$ is set to $1 \times 10^{-6}$ [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 $64 \times 64$.

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

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 (B1) 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× upampling is performed following previous studies. Blanked entries correspond to results not reported in previous works. Numbers in bold indicate the best performance.

<table>
<thead>
<tr>
<th>Model</th>
<th>Params (M)</th>
<th>Runtime (ms)</th>
<th>BI degradation</th>
<th>BD degradation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bicubic</td>
<td>-</td>
<td>-</td>
<td>26.14/0.7292</td>
<td>23.78/0.6347</td>
</tr>
<tr>
<td>VESPCN [1]</td>
<td>-</td>
<td>-</td>
<td>31.32/0.8684</td>
<td>25.35/0.7557</td>
</tr>
<tr>
<td>SPNC [49]</td>
<td>-</td>
<td>-</td>
<td>25.88/0.7752</td>
<td>-</td>
</tr>
<tr>
<td>TOFlow [61]</td>
<td>-</td>
<td>-</td>
<td>27.98/0.7990</td>
<td>25.89/0.7651</td>
</tr>
<tr>
<td>FRVSR [45]</td>
<td>5.1</td>
<td>137</td>
<td>-</td>
<td>37.09/0.922</td>
</tr>
<tr>
<td>DUF [21]</td>
<td>5.8</td>
<td>974</td>
<td>28.63/0.8251</td>
<td>-</td>
</tr>
<tr>
<td>RBPN [11]</td>
<td>12.2</td>
<td>1507</td>
<td>30.09/0.8590</td>
<td>37.12/0.8180</td>
</tr>
<tr>
<td>EDVR-M [57]</td>
<td>3.3</td>
<td>118</td>
<td>30.53/0.8699</td>
<td>27.10/0.8186</td>
</tr>
<tr>
<td>EDVR [57]</td>
<td>20.6</td>
<td>378</td>
<td>31.09/0.8800</td>
<td>27.35/0.8264</td>
</tr>
<tr>
<td>PFNL [65]</td>
<td>3.0</td>
<td>295</td>
<td>29.63/0.8502</td>
<td>26.73/0.8029</td>
</tr>
<tr>
<td>MuCAN [27]</td>
<td>-</td>
<td>30.88/0.8750</td>
<td>37.32/0.9465</td>
<td>-</td>
</tr>
<tr>
<td>TGA [18]</td>
<td>5.8</td>
<td>384</td>
<td>-</td>
<td>37.59/0.9156</td>
</tr>
<tr>
<td>RLSP [5]</td>
<td>4.2</td>
<td>49</td>
<td>-</td>
<td>38.48/0.9606</td>
</tr>
<tr>
<td>RSDN [17]</td>
<td>6.2</td>
<td>94</td>
<td>-</td>
<td>37.92/0.8505</td>
</tr>
<tr>
<td>RRN [19]</td>
<td>3.4</td>
<td>45</td>
<td>-</td>
<td>38.96/0.9444</td>
</tr>
<tr>
<td>BasicVSR [2]</td>
<td>6.3</td>
<td>63</td>
<td>31.42/0.8909</td>
<td>27.24/0.8251</td>
</tr>
<tr>
<td>IconVSR [2]</td>
<td>8.7</td>
<td>70</td>
<td>31.67/0.8948</td>
<td>27.39/0.8279</td>
</tr>
<tr>
<td>BasicVSR++ [4]</td>
<td>7.3</td>
<td>77</td>
<td>32.39/0.9069</td>
<td>27.79/0.8400</td>
</tr>
<tr>
<td>PSRT [47]</td>
<td>13.4</td>
<td>812</td>
<td>32.72/0.9106</td>
<td>28.07/0.8485</td>
</tr>
<tr>
<td>MFPI (Ours)</td>
<td>7.3</td>
<td>76</td>
<td>32.81/0.9106</td>
<td>28.11/0.8481</td>
</tr>
</tbody>
</table>

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.

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.
tions were performed on the REDS4 dataset [38] trained on the effectiveness of each component of MFPI. The evaluation performance benefits. As Table 2a suggests, combining all the method fails in VSR while our method leads to significant PSNR reduction, indicating the traditional model training to 0.12 PSNR improvements while KD leads to 0.38 dB improvements can be obtained by combining the two methods with the auxiliary branch. (b) Privilege training leads to 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.

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 $3 \times 3$, $7 \times 7$, $11 \times 11$, and $21 \times 21$. Besides, we only substituted the FFT with the original feature while keeping the LF, IN, and DWConv $21 \times 21$ (-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 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 $21 \times 21$ 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 $1 \times 1$ 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 5 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) MFE captures EFE with the energy function. (ii) Removing DCT from EFE leads to 0.43 dB PSNR drop, which indicates EFE is necessary. (iii) EFE with DWConv $1 \times 1$ kernel achieves the best PSNR than other DWConv kernels, indicating that a relatively larger kernel tends to achieve better performance.

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-

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

Figure 9: Visualization of intermediate features in MFPI.
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 DW-Conv. Additionally, we employ knowledge distillation [12] or privilege training (PT) to improve our results.

<table>
<thead>
<tr>
<th>Variant</th>
<th>PSNR</th>
</tr>
</thead>
<tbody>
<tr>
<td>Base + FFT</td>
<td>32.35</td>
</tr>
<tr>
<td>Base + FFT w/ Learnable filter</td>
<td>32.48</td>
</tr>
<tr>
<td>Base + FFT w/ Learnable filter + original feature</td>
<td>32.49</td>
</tr>
<tr>
<td>Base + FFT w/ Learnable filter + BN</td>
<td>32.43</td>
</tr>
<tr>
<td>Base + FFT w/ Learnable filter + IN</td>
<td>32.52</td>
</tr>
<tr>
<td>Base + FFT w/ Learnable filter + IN + DWConv 3 × 3</td>
<td>32.48</td>
</tr>
<tr>
<td>Base + FFT w/ Learnable filter + IN + DWConv 7 × 7</td>
<td>32.52</td>
</tr>
<tr>
<td>Base + FFT w/ Learnable filter + IN + DWConv 11 × 11</td>
<td>32.51</td>
</tr>
<tr>
<td>Base + FFT w/ Learnable filter + IN + DWConv 21 × 21 (Our SFE)</td>
<td>32.55</td>
</tr>
<tr>
<td>Base + w/ Learnable filter + IN + DWConv 21 × 21</td>
<td>31.93</td>
</tr>
</tbody>
</table>

(a) Individual components.

<table>
<thead>
<tr>
<th>Variant</th>
<th>PSNR</th>
</tr>
</thead>
<tbody>
<tr>
<td>Base + Energy function</td>
<td>32.40</td>
</tr>
<tr>
<td>Base + DCT w/ Fixed coefficients</td>
<td>32.32</td>
</tr>
<tr>
<td>Base + DCT w/ Learnable filter</td>
<td>32.41</td>
</tr>
<tr>
<td>Base + DCT w/ Fixed coefficients + Energy function</td>
<td>32.37</td>
</tr>
<tr>
<td>Base + DCT w/ Learnable filter + Energy function</td>
<td>32.50</td>
</tr>
<tr>
<td>Base + DCT w/ Learnable filter + Energy function + DWConv 3 × 3</td>
<td>32.42</td>
</tr>
<tr>
<td>Base + DCT w/ Learnable filter + Energy function + DWConv 7 × 7</td>
<td>32.40</td>
</tr>
<tr>
<td>Base + DCT w/ Learnable filter + Energy function + DWConv 11 × 11 (Our EFE)</td>
<td>32.52</td>
</tr>
<tr>
<td>Base + DCT w/ Learnable filter + Energy function + DWConv 21 × 21</td>
<td>32.45</td>
</tr>
<tr>
<td>Base + w/ Learnable filter + Energy function + DWConv 11 × 11</td>
<td>32.07</td>
</tr>
</tbody>
</table>

(b) Effects of the SFE branch.

<table>
<thead>
<tr>
<th>Variant</th>
<th>PSNR</th>
</tr>
</thead>
<tbody>
<tr>
<td>Base + DCT w/ Learnable filter + Energy function + DWConv 3 × 3</td>
<td>32.42</td>
</tr>
<tr>
<td>Base + DCT w/ Learnable filter + Energy function + DWConv 7 × 7</td>
<td>32.40</td>
</tr>
<tr>
<td>Base + DCT w/ Learnable filter + Energy function + DWConv 11 × 11 (Our EFE)</td>
<td>32.52</td>
</tr>
<tr>
<td>Base + DCT w/ Learnable filter + Energy function + DWConv 21 × 21</td>
<td>32.45</td>
</tr>
<tr>
<td>Base + w/ Learnable filter + Energy function + DWConv 11 × 11</td>
<td>32.07</td>
</tr>
</tbody>
</table>

(c) Effects of the EFE branch.

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

<table>
<thead>
<tr>
<th>Method</th>
<th>Needs Teacher?</th>
<th>PSNR</th>
</tr>
</thead>
<tbody>
<tr>
<td>BasicVSR</td>
<td>✓</td>
<td>31.58</td>
</tr>
<tr>
<td>+ Deep Supervision [23]</td>
<td>×</td>
<td>31.58</td>
</tr>
<tr>
<td>+ Self-KD [67]</td>
<td>×</td>
<td>31.47</td>
</tr>
<tr>
<td>+ Hinton KD [12]</td>
<td>✓</td>
<td>31.26</td>
</tr>
<tr>
<td>+ FitNet [44]</td>
<td>✓</td>
<td>31.52</td>
</tr>
<tr>
<td>+ PISR [24]</td>
<td>✓</td>
<td>31.62</td>
</tr>
<tr>
<td>+ Ours</td>
<td>×</td>
<td>31.76</td>
</tr>
</tbody>
</table>

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.

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


[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. 2


[50] Yapeng Tian, Yulun Zhang, Yun Fu, and Chenliang Xu Tdan. Temporally-deformable alignment network for video super-


