DFVSR: Directional Frequency Video Super-Resolution via Asymmetric and Enhancement Alignment Network

Shuting Dong[1,2], **Feng Lu**[1,2], **Zhe Wu**[2] and **Chun Yuan**[1,2] *

[1]Tsinghua Shenzhen International Graduate School, Tsinghua University
[2]Peng Cheng Laboratory
{dst21, lf22}@mails.tsinghua.edu.cn, wuzh02@pcl.ac.cn, yuanc@sz.tsinghua.edu.cn

Abstract

Recently, techniques utilizing frequency-based methods have gained significant attention, as they exhibit exceptional restoration capabilities for detail and structure in video super-resolution tasks. However, most of these frequency-based methods mainly have three major limitations: ***1)*** insufficient exploration of object motion information, ***2)*** inadequate enhancement for high-fidelity regions, and ***3)*** loss of spatial information during convolution. In this paper, we propose a novel network, Directional Frequency Video Super-Resolution (DFVSR), to address these limitations. Specifically, we reconsider object motion from a new perspective and propose Directional Frequency Representation (DFR), which not only borrows the property of frequency representation of detail and structure information but also contains the direction information of the object motion that is extremely significant in videos. Based on this representation, we propose a Directional Frequency-Enhanced Alignment (DFEA) to use double enhancements of task-related information for ensuring the retention of high-fidelity frequency regions to generate the high-quality alignment feature. Furthermore, we design a novel Asymmetrical U-shaped network architecture to progressively fuse these alignment features and output the final output. This architecture enables the intercommunication of the same level of resolution in the encoder and decoder to achieve the supplement of spatial information. Powered by the above designs, our method achieves superior performance over state-of-the-art models on both quantitative and qualitative evaluations.

1 Introduction

Video super-resolution (VSR) serves as an essential function in video processing. It solves how to reconstruct a high-resolution (HR) video from its corresponding low-resolution (LR) video. VSR has attracted considerable attention in both research and industrial communities because of its great value in video surveillance and autonomous vehicles.

Deep neural networks [Xue *et al.*, 2019; Wang *et al.*, 2019; Cao *et al.*, 2021; Chan *et al.*, 2021a; Chan *et al.*, 2021a; Liu *et al.*, 2022; Chan *et al.*, 2022], have made a substantial impact on enhancing the performance of VSR tasks. BasicVSR [Chan *et al.*, 2021a] designed a succinct pipeline to untangle the VSR task to four basic functionalities, i.e., Propagation, Alignment, Aggregation, and Upsampling. BasicVSR++ [Chan *et al.*, 2022] redesigned BasicVSR by proposing second-order grid propagation and flow-guided deformable alignment. VSR-T [Cao *et al.*, 2021] proposed a transformer-based method to exploit the locality and spatial-temporal data information through different layers to improve performance. These methods achieve remarkable performance on the VSR task. Nevertheless, they still face challenges in preserving sufficient details and avoiding structural artifacts. Recently, frequency-based methods [Xiao *et al.*, 2021; Qiu *et al.*, 2022; Leng *et al.*, 2022; Liu *et al.*, 2021; Fuoli *et al.*, 2019] have attracted great interest owing to their remarkable ability to recover details and structures in video restoration tasks. Natural images can be decomposed into high spatial frequency components that represent the image rapidly changing details and low spatial frequency components that describe the smoothly changing structure. Therefore, effective utilization of this property offers the possibility to recover details and structures efficiently.

Existing frequency-based methods [Xiao *et al.*, 2021; Qiu *et al.*, 2022; Leng *et al.*, 2022; Liu *et al.*, 2021; Fuoli *et al.*, 2019] developed a number of network structures, emphasizing the processing of frequency features to optimize the details and structure of video restoration. STD [Xiao *et al.*, 2021] derived spatial attention maps that indicated high-frequency video content. These maps were then employed to facilitate the transfer of spatial modeling between the networks, enhancing the overall performance of the process. FTVSR [Qiu *et al.*, 2022] and ICNet [Leng *et al.*, 2022] separated different frequencies to help improve the performance of VSR tasks. They have brought considerable performance gains. However, these methods have not fully explored the role of frequency on VSR tasks for the following reasons. **Firstly**, previous frequency-based methods mainly divide images into high- and low-frequency or different levels of frequency representations. However, such frequency-

*Corresponding Author

based representations only describe the appearance information (including structure and texture details) of the object. But, objects in video restoration tasks are directional in motion. These representations ignore the motion direction information of the object. **Secondly**, not all frequency information is equally useful for video restoration tasks. Without selecting and enhancing effective information, the frequency-based model may easily attend to many low-fidelity and less informative frequency regions. This may degrade model efficiency. **Thirdly**, existing frequency-based works commonly adopted down-sampling to enlarge the receptive field and extract global information, resulting in the inevitable loss of spatial information. However, VSR tasks are pixel-wise dense problems. Sharp predictions would not be made without the assistance of sufficient spatial information.

In this paper, we propose a novel VSR network, DFVSR, to cope with the above problems. **Firstly**, we observe the object motion as the combination of movements in different directions information (horizontal, vertical, and diagonal). This coincides with the fact that natural images (frames) can be decomposed into multiple high and low frequencies in different directions. Inspired by this observation, we reconsider motion from a new perspective and propose to view the motion as representations of frequencies in different directions. This way effectively borrows the property of frequency representation of texture details and structure information, and importantly, it also contains motion information in multiple directions that is extremely significant in videos. **Secondly**, as shown in Figure 2, based on the above representation, we propose a Directional Frequency-Enhanced Alignment (DFEA) for directing the alignment to pay attention to the regions with high-fidelity frequency representations to generate high-quality alignment features. We design DFEA to use double enhancements of task-related information to ensure the retention of valid information and the weakening of invalid information. Meanwhile, DFEA also alleviates the limitation of DCN instability during training. **Finally**, we design a novel Asymmetrical U-shaped network architecture to progressively fuse the alignment features. This architecture enables the intercommunication of the same level of resolution in the encoder and decoder to achieve the supplement of spatial information. Powered by the above designs, we surpass the state-of-the-art (SOTA) methods in the VSR task.

Our contributions can be summarized as follows:

- We propose a novel network, DFVSR, for improving the performance of VSR tasks. To the best of our knowledge, it is the first attempt to explore the potential of directional frequency in VSR tasks.
- We present a novel implicit alignment, DFEA, to pay attention to the regions with high-fidelity frequency representations to generate high-quality alignment features. DFEA also alleviates the limitation of DCN instability during training.
- We customize a novel Asymmetrical U-shaped architecture to enable the intercommunication of the same level of resolution in the encoder and decoder to achieve the supplement of spatial information.
- Experimental results demonstrate the superiority of our model over SOTA models on both quantitative and qualitative evaluations.

2 Related Work

Frequency-Based Methods. Recently, frequency-based methods [Xiao *et al.*, 2021; Qiu *et al.*, 2022; Leng *et al.*, 2022; Liu *et al.*, 2021; Fuoli *et al.*, 2019] have developed a variety of network architectures that concentrate on processing frequency features to enhance the details and structure of image and video restoration. STD [Xiao *et al.*, 2021] extracted spatial attention maps, which represented high-frequency video content from both networks. These maps were subsequently utilized to streamline the transfer of spatial modeling between the networks, ultimately improving the overall performance of the process. Additionally, methods like FTVSR [Qiu *et al.*, 2022] and ICNet [Leng *et al.*, 2022] separated various frequencies to effectively boost the performance of VSR tasks. These frequency-based methods have resulted in significant performance improvements. However, such frequency-based representations ignore the motion information of the object. In this paper, we propose a more effective representation, DFR, to simultaneously model detail and structure information, as well as motion information.

Alignment Methods. The existing alignment works are mainly divided into two categories: (i) flow-based alignment that exploits optical flow to predict motion fields, and (ii) deformable convolution (DCN) based alignment that employs DCN [Dai *et al.*, 2017] to perform implicit feature learning. The flow-based alignment methods rely heavily on flow estimation, and any errors in the flow computation may introduce artifacts around frame structures [Chan *et al.*, 2021a; Tian *et al.*, 2020]. DCN-based alignment methods [Chan *et al.*, 2021a; Chan *et al.*, 2021b; Zhu *et al.*, 2019] have demonstrated significant improvements over flow-based alignment thanks to the diversity of its offset. Nonetheless, DCN-based alignment can be difficult to train [Chan *et al.*, 2021b]. BasicVSR++ [Chan *et al.*, 2022] proposed a flow-guided deformable alignment to help offset learning by using optical flow field guidance. Unlike the aforementioned methods, we propose to pay attention to the regions with high-fidelity frequency representations to generate more valid features. Then, these features are fed to implicit alignment. This way alleviates the limitation of DCN instability during training.

3 Method

3.1 Overall Architecture

The overall architecture of DFVSR is shown in Figure 1 (a). We propose a novel directional frequency representation in DFVSR, which not only borrows the property of frequency representation of detail and structure information but also contains the direction information of the object motion that is extremely significant in videos. DFVSR consists of multiple units, and each unit is abbreviated as DFunit (as shown in Figure1 (b)). We design the DFunit with a directional frequency guidance encoder, a bottleneck, and an asymmetrical decoder. In the encoder of DFUnit, the initial input is low-frequency features, which are progressively fused with

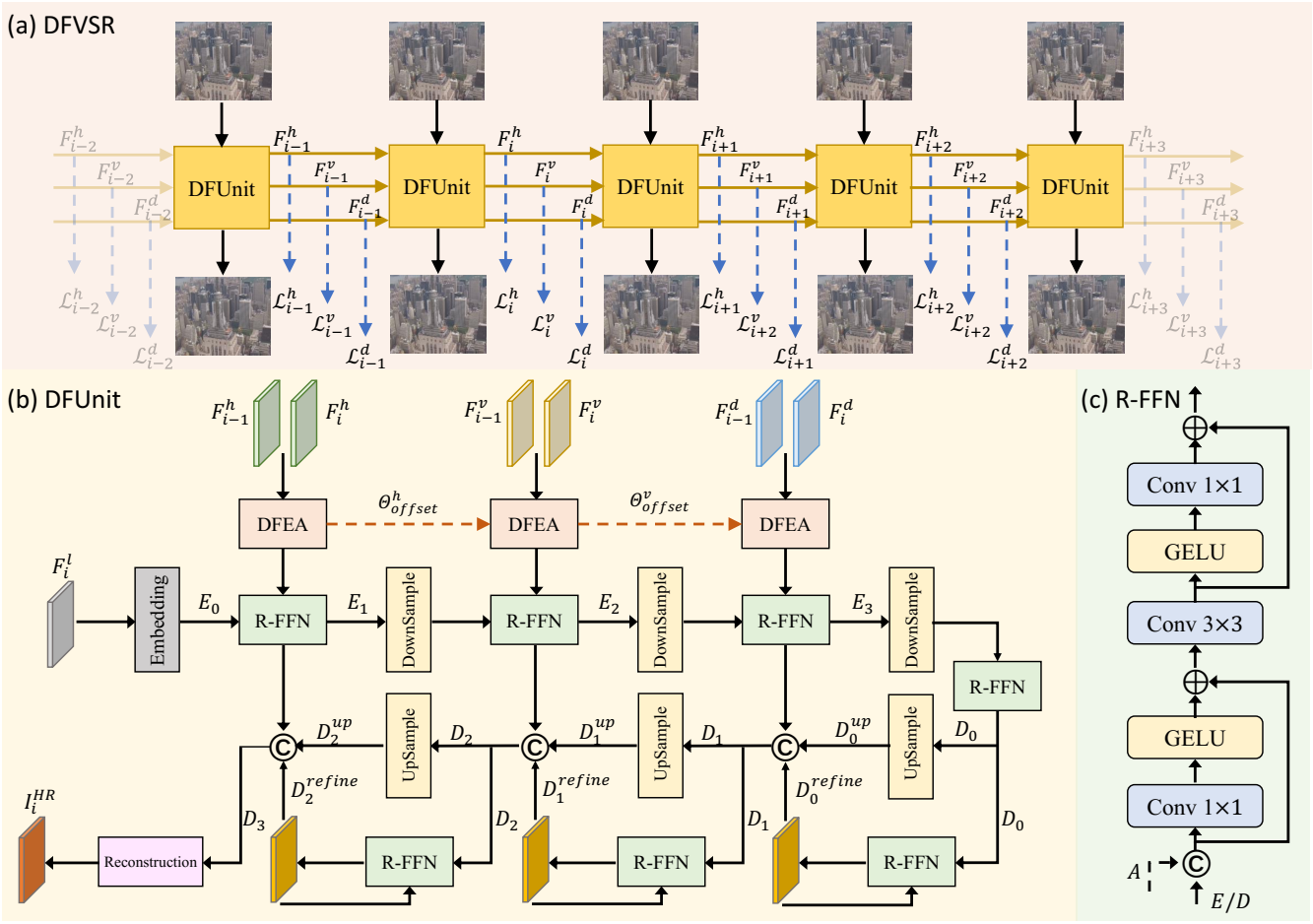

Figure 1: (a) The overall architecture of DFVSR. (b) Illustration of DFVSR Unit (DFUnit). DFUnit adopts an Asymmetric U-shaped structure that consists of a directional frequency guidance encoder, a bottleneck, and an asymmetrical decoder. (c) The architecture of R-FFN.

higher-frequency alignment features. We propose a novel implicit alignment, DFEA, to generate these alignment features (as shown in Figure 2). DFEA uses double enhancements of task-related information to ensure the retention of high-fidelity frequency regions to generate high-quality alignment features. Then, the output of the DFUnit encoder is fed into an asymmetric decoder through a bottleneck. This decoder further processes, refines, and reconstructs features to output the final HR frame. The above design ensures the final video is a high-quality product.

3.2 Directional Frequency Representation

For high-quality output, the restoration of rich texture details and clear structure is essential. Natural images can be decomposed into high spatial frequency components that represent rapidly changing texture details of images and low spatial frequency components that describe the smoothly changing structure. Thus, efficient exploration of this property opens up possibilities for high-quality restoration of texture details and structure. It brings considerable performance gains for frequency-based methods on video restoration tasks. Previous frequency-based methods mainly divide images into high- and low-frequency maps or different levels of frequency. However, such frequency representations do not contain object motion information, ignoring the directionality of object motion.

To handle this limitation, it is necessary to develop a directional representation for precisely detecting the motion directions of objects while providing detail and structure information. We propose Directional Frequency Representation (DFR), which not only borrows the property of frequency representation of detail and structure information but also contains the direction information of the object motion that is extremely significant in videos. Specifically, the i^{th} frame feature G_i can be represented as:

$$G_i = \sum_j F_i^j, j \in \{l, h, v, d\}, \tag{1}$$

$$F_i^j = \gamma(H_i, j), j \in \{l, h, v, d\}, \tag{2}$$

where F denotes the frequency feature, and j reflects the frequency direction. l represents low frequency. h, v, and d

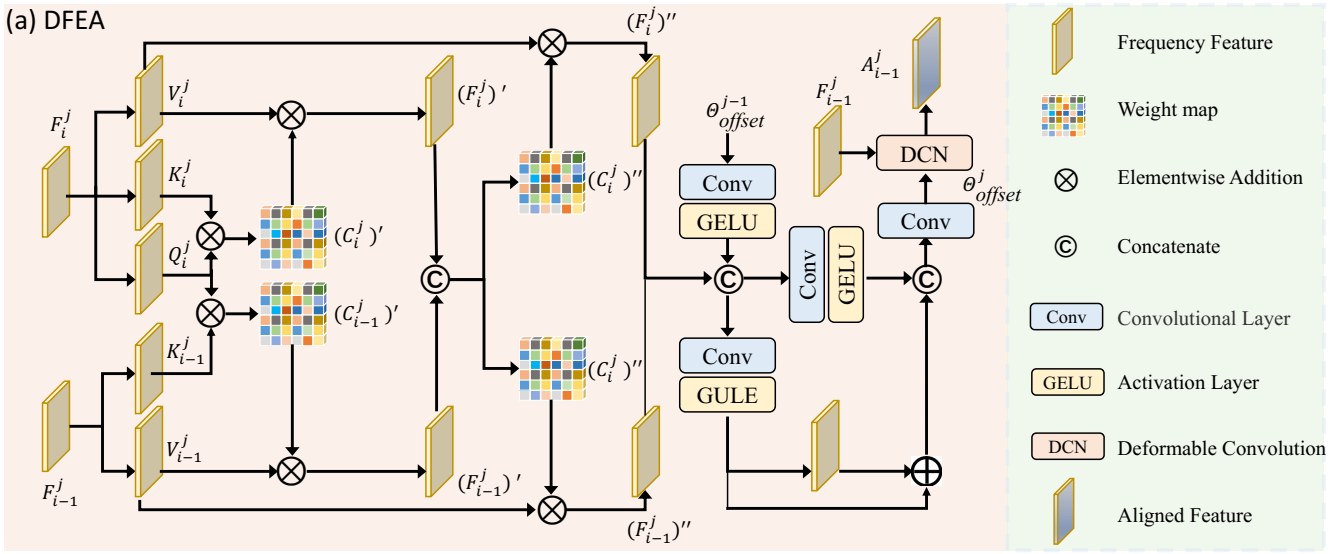

Figure 2: Overview of our proposed DFEA module. DFEA uses double enhancements of task-related information to ensure the retention of high-fidelity frequency regions.

represent the high frequencies in the horizontal, vertical, and diagonal directions, respectively. γ is the discrete wavelet transform operation for generating the directional frequency features.

Different from previous frequency-based methods, this representation incorporates directional information, thus object motion can be viewed as an amalgamation of frequency representations from multiple directions. This will provide more sufficient motion information for subsequent inter-frame alignment.

Directional Frequency Loss. We argue that the directional information of motion in videos is significant for subsequent inter-frame motion alignment operations. Unlike images, an important property of objects in video is their motion. As analyzed above, in previous frequency-based methods, the motion information of objects may be not well modeled. It will cause great challenges for feature alignment and texture refinement. Therefore, we hope our model could not only focus on the appearance (texture details and structure) information of objects but also could pay more attention to the object motions. As shown in Figure 1(a), we propose a novel Directional Frequency Loss (DFLoss) to ensure the accuracy of motion information in the propagation process. We define the DFLoss:

$$\mathcal{L}_i^j = \sum_j \left\| \varphi\left(F_i^j\right) - \varphi(\mathcal{E}\left(F_i^{j-1}, F_i^j\right)) \right\|_2^2, j \in \{l, h, v, d\}, \tag{3}$$

where $\mathcal{E}$ is the proposed DFEA (alignment module) that will be described in detail in section 3.4. φ represents the reconstruction layer. In this way, directional frequency maps act as the appearance and motion guidance for the objects in videos. The proposed DFLoss enforces the ingredients of appearance and motion in deep feature layers for blending. Through propagating the directional frequency features, the motion information will be effectively retained.

3.3 Asymmetric Encoder and Decoder

Directional Frequency Guidance Encoder. Different from the previous methods mainly based on using different branches to separately process multiple frequency features, we propose a novel encoder that progressively incorporates alignment frequency features.

In our design, the computationally intensive alignment operation is assigned to high-frequency features. This strategic allocation of network parameters shifts more attention towards detail-representing high-frequency features, while significantly reducing the computational load associated with low-frequency features. As shown in Figure 1(b), we first input the low-frequency feature of i^{th} frame and progressively feed the alignment directional high-frequency features. Specifically, the low frequency F_i^l is further extracted and generates the feature E_0. E_0 is the input of the encoder, which consists of several R-FFNs (as shown in Figure 1(c)). We propose an R-FFN module to improve the feed-forward network (FFN) by adding residual connections. Each encoder block contains two inputs: the output features derived from the downsampling operation applied to the previous encoder block, and the corresponding directional alignment frequency feature generated by our proposed DFEA module. The output E_m of m^{th} encoder block is defined as:

$$E_m = C(E_{m-1}, \mathcal{E}(F_{i-1}^j, F_i^j)) \tag{4}$$

where $\mathcal{E}$ is the proposed DFEA (alignment module) that will be described in detail in section 3.4. C denotes the R-FFNs. The output of each encoder block will be sent to two places for two different purposes. One is to conduct a downsampling operation before feeding into the next encoder block. The other is performed as an added skip connection to provide supplementary spatial information to the decoder.

After this encoding process, the final output feature E_3 of the encoder, incorporates alignment frequency information in multiple directions. E_3 then passes through the bottleneck, which consists of an R-FFN module to output the feature D_0.

Asymmetrical Decoder. We design a novel asymmetrical decoder that is able to complement the lost spatial information. The n^{th} decoder block includes two inputs: 1) the output D_{n-1} of the previous decoder block, and 2) the output E_m of the corresponding m^{th} encoder block. These two inputs are not simultaneously but sequentially input to the decoder block. D_{n-1} is input to the upsampling part and R-FFN, respectively. We design the features generated by R-FFN to be fed back into R-FFN iteratively to refine the encoder feature. We design the iteration parameter as the hyperparameter (set to 3 in this paper). Then, we concatenate the feature D_n^{up} generated by upsampling part, the feature D_n^{refine} generated by the R-FFN iterative refinement part, and the feature E_m to generate the output D_n of n^{th} decoder block. The decoder process is defined as:

$$D_n = Concat(E_m, D_n^{up}, D_n^{refine}). \tag{5}$$

The output feature D_3 of the last decoder block is then fed into the reconstruction layer to output the i^{th} video frame I_i^{HR}:

$$I_i^{HR} = R(D_3) \tag{6}$$

where R is the reconstruction layer.

3.4 Directional Frequency-Enhanced Alignment

The other limitation of previous frequency-based methods is that the learning model may attend to some less informative spatial regions with low-fidelity frequency representations. This may degrade model efficiency. Ideally, the model would be able to pay more attention to spatial regions with higher fidelity information. This goal motivates us to design an alignment module that focuses more on high-fidelity regions. Therefore, we propose a novel alignment model, DFEA, which consists of double enhancements of task-related information to ensure the retention of valid information in high-fidelity regions and the weakening of invalid information in low-fidelity regions.

As shown in Figure 2, we map F_i^j into the query Q_i^j, the key K_i^j, and the value V_i^j. Differently, we map F_{i-1}^j into the key K_{i-1}^j and the value V_{i-1}^j. We do this to preserve task-relevant information in i^{th} frame, as well as adaptively select alignment information in $(i-1)^{th}$ frame. Specifically, we first utilize Q_i^j, K_i^j and K_{i-1}^j to calculate the correlation matrix $(C_i^j)'$ and $(C_{i-1}^j)'$. Next, the correlation matrix multiplies its corresponding value to obtain the correlation features $\left(F_i^j\right)'$ and $\left(F_{i-1}^j\right)'$. Then, to further enhance the correlation information, we concatenate $\left(F_i^j\right)'$ and $\left(F_{i-1}^j\right)'$ to calculate the enhanced matrix $(C_i^j)''$ and $(C_{i-1}^j)''$. Subsequently, the enhanced task-relevant features $\left(F_i^j\right)''$ and $\left(F_{i-1}^j\right)''$ are generated by the matrix multiplication between the enhanced matrices and the original frequency value features.

In addition, we argue that the frequencies of different directions for the same frame are not independent but correlated. Therefore, the offsets obtained by learning different direction frequencies in the same frame should also be related to each other. Based on this analysis, as shown in Figure 2, we design to use previously learned offers Θ_{offset}^{j-1} to help learn offers Θ_{offset}^{j} in the current direction. Here, $j-1$ does not refer to the mathematical meaning, but the previous frequency direction. Last, the alignment features A_{i-1}^j of the j^{th} direction is generated as follows:

$$A_{i-1}^j = DC(F_{i-1}^j, \Theta_{offset}^j) \tag{7}$$

where $DC(\cdot)$ denotes deformable convolution.

Unlike the original DCN-based alignment to directly concatenate two features to learn offsets, our proposed DFEA enhances the valid information of i^{th} and $i-1^{th}$ frames and ensures that its learning offsets are more accurate and valid. This way makes the alignment process more stable.

4 Experiments

4.1 Datasets and Implementation

Datasets. We adopt two widely used datasets to train: REDS [Nah *et al.*, 2019] and Vimeo-90K [Xue *et al.*, 2019]. Following [Chan *et al.*, 2021a], we apply REDS4 as our test set, and REDSval4 as the validation set. We also use Vid4 [Liu and Sun, 2013], UDM10 [Yi *et al.*, 2019] and Vimeo-T [Xue *et al.*, 2019] as test sets.

Implementation and Training Details. We employ Adam optimizer by setting $\beta_1 = 0.9$ and $\beta_2 = 0.999$. The learning rate is initialized as 2.5×10^{-5}. We apply RGB patches of size 64×64 as inputs. We set the mini-batch size to 32. In addition to our proposed DFLoss, we also adopt Charbonnier loss [Lai *et al.*, 2017], and ε is set to 1×10^{-3}. The total number of iterations is 600K. The model is trained under the PyTorch framework with an NVIDIA RTX 2080Ti GPU. To speed up the convergence, we develop shallower networks to initialize deeper network parameters.

4.2 Comparison with State-of-The-Art Methods

We compare our method with 12 SOTA models that are BasicVSR++ [Chan *et al.*, 2022], TTVSR [Liu *et al.*, 2022], IconVSR [Chan *et al.*, 2021a], BasicVSR [Chan *et al.*, 2021a], VSR-T [Cao *et al.*, 2021], TDAN[Tian *et al.*, 2020], EDVR [Wang *et al.*, 2019], TOFlow [Xue *et al.*, 2019], FRVSR [Sajjadi *et al.*, 2018], DUF [Jo *et al.*, 2018], RBPN [Haris *et al.*, 2019] and MuCAN [Huang *et al.*, 2017]. The results are in Table 1. Our DFVSR achieves SOTA performance on all datasets for both BI and BD degradations. Compared to BasicVSR++ [Chan *et al.*, 2022] which is the representative progressive method, our model achieves superior performance on all datasets by using fewer parameters. Furthermore, DFVSR also surpasses the recent transformer-based method VSR-T [Cao *et al.*, 2021] by 1.57 dB, 1.54 dB, and 0.56 dB, respectively. The qualitative comparisons are shown in Figure 3. These visual results show that our method restores clearer details and a more precise structure. Furthermore, our method is also very efficient in the recovery of bright and dark areas. This gain is considered significant for VSR tasks.

		BI degradation			BD degradation		
	Params (M)	REDS4	Vimeo-T	Vid4	UDM10	Vimeo-T	Vid4
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
TOFlow	–	27.98/0.7990	33.08/0.9054	25.89/0.7651	36.26/0.9438	34.62/0.9212	25.85/0.7659
FRVSR	5.1	37.09/0.9522	35.64/0.9319	26.69/0.8103			
DUF	5.8	28.63/0.8251	–	–	38.48/0.9605	36.87/0.9447	27.38/0.8329
RBPN	12.2	30.09/0.8590	37.07/0.9435	27.12/0.8180	38.66/0.9596	37.20/0.9458	–
EDVR	20.6	31.09/0.8800	37.61/0.9489	27.35/0.8264	39.89/0.9686	37.81/0.9523	27.85/0.8503
MuCAN	–	30.88/0.8750	37.32/0.9465	–	–	–	–
VSR-T	32.6	31.19/0.8815	37.71/0.9494	27.36/0.8258	–	–	–
BasicVSR	6.3	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	8.7	31.67/0.8948	37.47/0.9476	27.39/0.8279	40.03/0.9694	37.84/0.9524	28.04/0.8570
TTVSR	6.8	32.12/0.9021	–	–	40.41/0.9712	37.92/0.9526	28.40/0.8643
BasicVSR++	7.3	32.39/0.9069	37.79/0.9500	27.79/0.8400	40.72/0.9722	38.21/0.9550	29.04/0.8753
DFVSR (Ours)	7.1	32.76/0.9081	38.25/0.9556	27.92/0.8427	40.97/0.9733	38.51/0.9571	29.56/0.8983

Table 1: Quantitative comparison (PSNR and SSIM) of different methods on REDS4, Vimeo-T, Vid4 and UDM10 with upscale factor 4 under BI and BD degradations. The top-2 results are highlighted in red and blue colors.

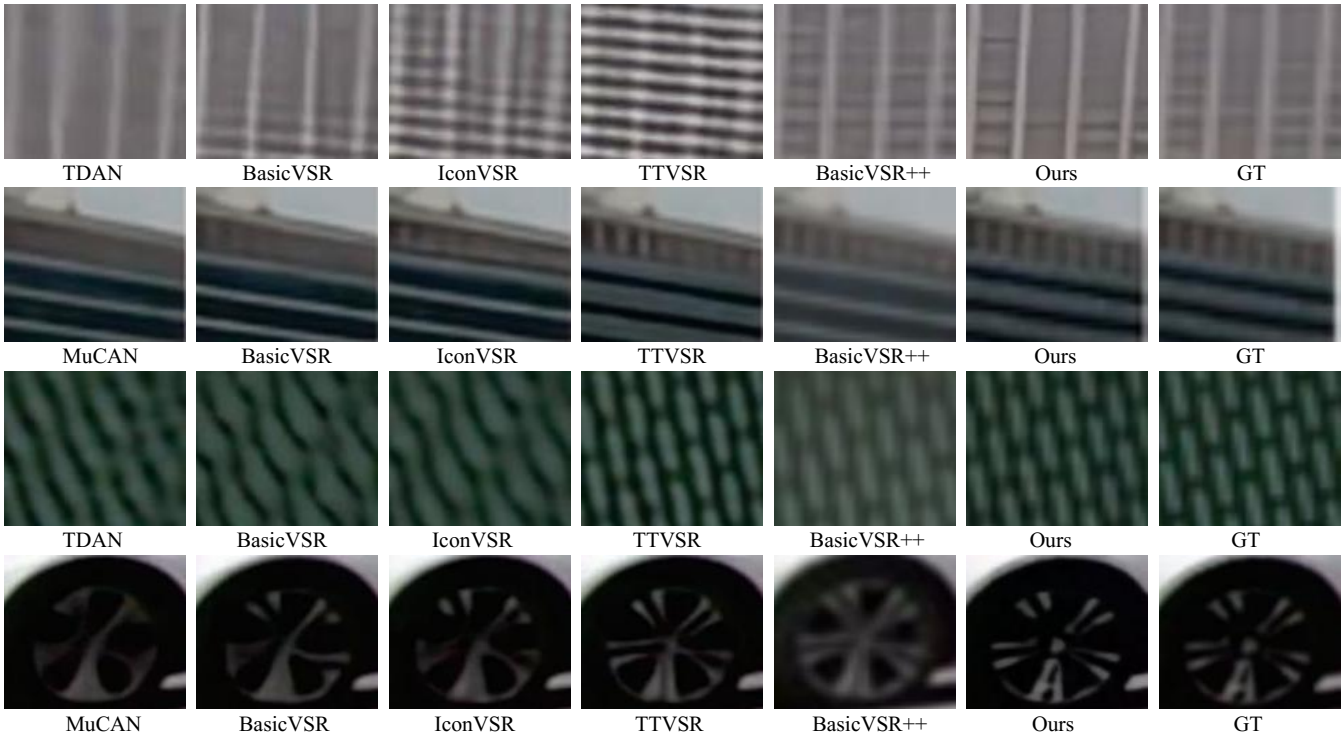

Figure 3: Challenging scenario. The results demonstrate that our method has the ability to restore more details and clearer structure.

4.3 Ablation Study

We conduct the ablation study to explore the contributions of the proposed components. We start with a baseline, which is the cascaded R-FFNs. We denote the model generated by the baseline as A_0. A_0 uses original non-frequency features to train. Then, we gradually insert the components to validate the necessity of each component. The quantitative results are in Table 2.

Effectiveness of the input strategy of frequency features. There are two strategies for input here. One is to input different frequencies into the model training together, and the other is to gradually input different frequencies into the model training. We train the two strategies separately. The model generated by the former strategy is denoted as A_{1-1}, and the latter strategy applies the proposed model to gradually incorporates frequency features to generate a model, which is denoted as A_{1-2}. The results show that A_{1-2} is better. This illustrates that progressively fusing the frequencies yields better results.

Effectiveness of the DFLoss module. We test the contribution of our proposed DFLoss based on the A_{1-2} model (trained only by Charbonnier loss). We design to train two models (A_{2-1} and A_{2-2}), which use Charbonnier loss $\mathcal{L}_{CL}$

	A_0	A_{1-1}	A_{1-2}	A_{2-1}	A_{2-2}	A_{3-1}	A_{3-2}	A_{4-1}	A_{4-2}	A_{4-3}
Input strategy	✗	✗	✓	✓	✓	✓	✓	✓	✓	✓
DFLoss	✗	✗	✗	✓	✓	✓	✓	✓	✓	✓
Asymmetrical U-shape	✗	✗	✗	✗	✗	✗	✓	✓	✓	✓
DFEA	✗	✗	✗	✗	✗	✗	✗	✗	✗	✓
PSNR	32.27	33.02	34.79	35.33	35.62	35.91	36.73	36.81	37.94	38.51

Table 2: Ablation studies on Vimeo-T with upscale factor 4 under BD degradation. A in the table represents different models.

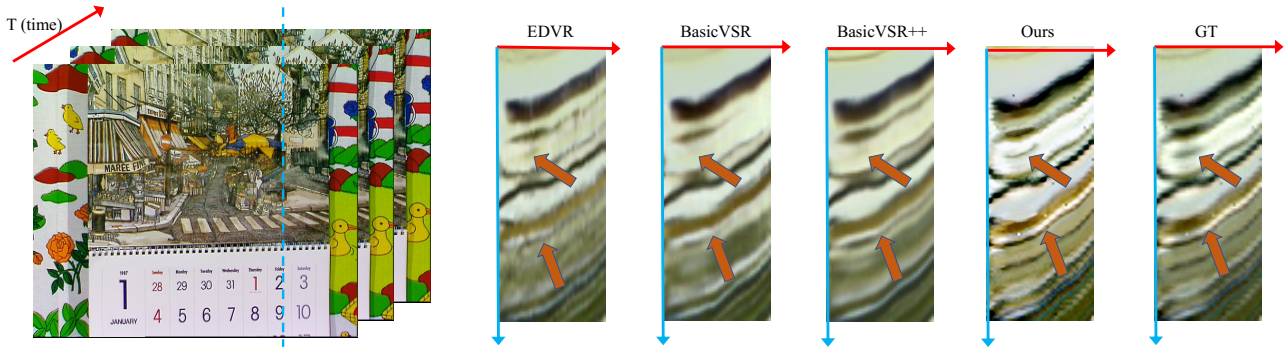

Figure 4: Temporal consistency comparison. Only our method recovers the details of the area pointed by the orange arrow. Also, our method recovers more details in the overall profile to make the output more temporally consistent.

and DFLoss $\mathcal{L}_{DF}$. The total loss is defined as: $\mathcal{L} = \lambda_1 \mathcal{L}_{CL} + \lambda_2 \mathcal{L}_{DF}$. To train A_{2-1}, we set $\lambda_1 = 0.7$ and $\lambda_2 = 0.3$. On the contrary, to train A_{2-2}, we set $\lambda_1 = 0.3$ and $\lambda_2 = 0.7$. The experimental results show that both A_{2-1} and A_{2-2} models are better than A_{1-2}. And these two models gain improvements of 0.54dB and 0.83dB respectively. In addition, the performance of A_{2-2} is further better. This shows that directional frequencies can be an effective guide for the appearance and motion of objects in video.

Effectiveness of the Asymmetrical U-shaped architecture. To explore the contributions of the proposed Asymmetrical U-shaped architecture. We train two U-shape models based on the A_{2-2} model. A model trained using a common symmetric U-shape architecture (ie, delete D_{refine} for Equation (5)), called A_{3-1}. Another model is trained with our proposed asymmetric architecture, called A_{3-2}. As shown in Table 2, the results of both A_{3-1} and A_{3-2} are better than A_{2-2}. This shows the importance of supplementary spatial information. In addition, the proposed A_{3-2} achieves better results. It demonstrates the superiority of the proposed Asymmetric U-shape. It also shows that our proposed architecture can supplement more effective spatial information. This shows the necessity of further refinement of the decoder features.

Effectiveness of the DFEA module. We designed three alignment modules based on A_{3-2}. The first alignment module, that is, the original way of DCN-based alignment, performs a concatenate operation on the frequency features of the input two frames. The second alignment module uses one attention operation on the frequency features of the input two frames. The third alignment module is our proposed DFEA, which designs double enhancements of task-related information to ensure the retention of high-fidelity frequency regions. The models generated by these three modules are respectively denoted as A_{4-1}, A_{4-2}, and A_{4-3}. The experimental results show that A_{4-3} (DFVSR) has the best performance, and A_{4-2} is the second best. This reflects the importance of enhancing the weight of high-fidelity regions. In addition, it demonstrates the efficiency of our alignment method.

4.4 Temporal Consistency

In Figure 4, we show the comparisons of the temporal profiles between our DFVSR and three SOTA VSR methods, BasicVSR++ [Chan *et al.*, 2022], BasicVSR [Chan *et al.*, 2021a] and EDVR [Wang *et al.*, 2019]. We collect a column (blue dotted lines) and obtain the temporal profile to compare temporal consistency. We can observe that our DFVSR produces smoother temporally consistent results by comparing all of the methods. Furthermore, in the area indicated by the orange arrow, the comparison results demonstrate that our model preserves more details and produces high-quality products with temporal consistency.

5 Conclusion

In this paper, we propose a novel and effective DFVSR network to improve the performance of VSR tasks. We propose a novel representation, DFR, which not only borrows the property of frequency representation of detail and structure information but also contains the object motion information that is extremely significant in videos. Based on this representation, we propose a new implicit alignment module, DFEA, to ensure the retention of high-fidelity frequency regions to generate the high-quality alignment feature. Furthermore, we design a novel Asymmetrical U-shaped network architecture to progressively fuse these alignment features and output the final output. This architecture enables the intercommunication of the same level of resolution in the encoder and decoder to achieve the supplement of spatial information. The above designs ensure the final video is a high-quality product.

Acknowledgments

This work was supported by the National Key R&D Program of China (2022YFB4701400/4701402), SZSTC Grant (JCYJ20190809172201639, WDZC20200820200655001), Shenzhen Key Laboratory (ZDSYS20210623092001004). This work was also supported by National Natural Sciences Foundation of China (62102207).

References

[Cao *et al.*, 2021] Jiezhang Cao, Yawei Li, Kai Zhang, and Luc Van Gool. Video super-resolution transformer. *arXiv preprint arXiv:2106.06847*, 2021.

[Chan *et al.*, 2021a] Kelvin CK Chan, Xintao Wang, Ke Yu, Chao Dong, and Chen Change Loy. Basicvsr: The search for essential components in video super-resolution and beyond. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4947–4956, 2021.

[Chan *et al.*, 2021b] 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.

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

[Dai *et al.*, 2017] Jifeng Dai, Haozhi Qi, Yuwen Xiong, Yi Li, Guodong Zhang, Han Hu, and Yichen Wei. Deformable convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, pages 764–773, 2017.

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

[Haris *et al.*, 2019] 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*, pages 3897–3906, 2019.

[Huang *et al.*, 2017] 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.

[Jo *et al.*, 2018] 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 conference on computer vision and pattern recognition*, pages 3224–3232, 2018.

[Lai *et al.*, 2017] Wei-Sheng Lai, Jia-Bin Huang, Narendra Ahuja, and Ming-Hsuan Yang. Deep laplacian pyramid networks for fast and accurate super-resolution. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 624–632, 2017.

[Leng *et al.*, 2022] Jiaxu Leng, Jia Wang, Xinbo Gao, Bo Hu, Ji Gan, and Chenqiang Gao. Icnet: Joint alignment and reconstruction via iterative collaboration for video super-resolution. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 6675–6684, 2022.

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

[Liu *et al.*, 2021] Shaoli Liu, Chengjian Zheng, Kaidi Lu, Si Gao, Ning Wang, Bofei Wang, Diankai Zhang, Xiaofeng Zhang, and Tianyu Xu. Evsrnet: Efficient video super-resolution with neural architecture search. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2480–2485, 2021.

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

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

[Qiu *et al.*, 2022] Zhongwei Qiu, Huan Yang, Jianlong Fu, and Dongmei Fu. Learning spatiotemporal frequency-transformer for compressed video super-resolution. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XVIII*, pages 257–273. Springer, 2022.

[Sajjadi *et al.*, 2018] Mehdi SM Sajjadi, Raviteja Vemulapalli, and Matthew Brown. Frame-recurrent video super-resolution. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6626–6634, 2018.

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

[Wang *et al.*, 2019] Xintao Wang, Kelvin CK Chan, Ke Yu, Chao Dong, and Chen Change Loy. Edvr: Video restoration with enhanced deformable convolutional networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 0–0, 2019.

[Xiao *et al.*, 2021] Zeyu Xiao, Xueyang Fu, Jie Huang, Zhen Cheng, and Zhiwei Xiong. Space-time distillation for

video super-resolution. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2113–2122, 2021.

[Xue *et al.*, 2019] 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.

[Yi *et al.*, 2019] Peng Yi, Zhongyuan Wang, Kui Jiang, Junjun Jiang, and Jiayi Ma. Progressive fusion video super-resolution network via exploiting non-local spatio-temporal correlations. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 3106–3115, 2019.

[Zhu *et al.*, 2019] Xizhou Zhu, Han Hu, Stephen Lin, and Jifeng Dai. Deformable convnets v2: More deformable, better results. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9308–9316, 2019.