Multi-Axis Feature Diversity Enhancement for Remote Sensing Video Super-Resolution

Yi Xiao, *Graduate Student Member, IEEE*, Qiangqiang Yuan, *Member, IEEE*, Kui Jiang, *Member, IEEE*, Yuzeng Chen, Shiqi Wang, *Senior Member, IEEE*, and Chia-Wen Lin, *Fellow, IEEE*

***Abstract*—How to aggregate spatial-temporal information plays an essential role in video super-resolution (VSR) tasks. Despite the remarkable success, existing methods adopt static convolution to encode spatial-temporal information, which lacks flexibility in aggregating information in large-scale remote sensing scenes, as they often contain heterogeneous features (*e.g.,* diverse textures). In this paper, we propose a spatial feature diversity enhancement module (SDE) and channel diversity enhancement module (CDE), which explore the diverse representation of different local patterns while aggregating the global response with compactly channel-wise embedding representation. Specifically, SDE introduces multiple learnable filters to extract representative spatial variants and encodes them to generate a dynamic kernel for enriched spatial representation. To explore the diversity in the channel dimension, CDE exploits the discrete cosine transform to transform the feature into the frequency domain. This enriches the channel representation while mitigating massive frequency loss caused by pooling operation. Based on SDE and CDE, we further devise a multi-axis feature diversity enhancement (MADE) module to harmonize the spatial, channel, and pixel-wise features for diverse feature fusion. These elaborate strategies form a novel network for satellite VSR, termed MADNet, which achieves favorable performance against state-of-the-art method BasicVSR++ in terms of average PSNR by 0.14 dB on various video satellites, including JiLin-1, Carbonite-2, SkySat-1, and UrtheCast. Code will be available at https://github.com/XY-boy/MADNet**

***Index Terms*—Video super-resolution, dynamic convolution, frequency analysis, remote sensing.**

I. Introduction

VIDEO satellite, one of the promising earth observation techniques, has recently attracted increasing attention due to its dynamic observation capability. Compared to traditional static remote sensing images, video satellites could record specific areas by adjusting the optical axis of satellite sensors, providing continuous observation for various applications, such as object tracking [1], classification [2], [3], segmentation [4], [5], [6], etc. However, videos captured from satellite platforms face complex degradation [7], [8], *e.g.,* platform tremors and scattering. Moreover, to stabilize the ultra remote transmission between satellite and ground station, satellite video often suffers from compression and downsampling, resulting in undesirable high-frequency information loss. To this end, it is essential to develop a practical scheme, like super-resolution (SR) technologies, to improve the spatial resolution of satellite videos.

SR is a long-standing ill-posed issue, which aims to reconstruct a high-resolution (HR) output from the given low-resolution (LR) observation [9], [10], [11], [12]. Compared to single-image super-resolution (SISR), VSR is more complicated as it requires aggregating both spatial and temporal information from misaligned video frames. Prior to deep learning, conventional model-based methods rely on specific assumptions and priors, *e.g.,* non-local mean [13] and Bayesian [14], to address complex motions. However, these handcrafted priors are of limited representation ability, especially for highly complex and varied motion scenarios.

With the great success of deep learning in a variety of areas [15], [16], [17], SR algorithms based on deep learning are studied extensively. Researchers promote further progress in innovative architectures and training practices for VSR and show considerable superiority over conventional algorithms in visual quality improvement [18], [19], [20], [21]. For example, prior works often exploit sliding-window fashion [18], [22], [23], [24] to explore useful redundancy information. Since the accessible information is limited in the local neighborhood, the potential complementarities in the global respective field have been barely explored. More recently, the recurrent network [20], [25], [26] has been proposed to aggregate spatial-temporal information sequentially. Among them, the bidirectional propagation scheme [19], [27], [28] has demonstrated impressive performance in simulating the temporal motion. They use globally-shared convolution to encode spatial-temporal context for feature propagation.

Nevertheless, while they have demonstrated favorable performance, some potential problems still exist in large-scale earth observation scenarios, making VSR more challenging. **Firstly, as shown in Fig. 2, there exists spatial diversity of**

Received 15 October 2024; revised 15 January 2025; accepted 26 February 2025. Date of publication 7 March 2025; date of current version 21 March 2025. This work was supported in part by the National Natural Science Foundation of China under Grant 423B2104, Grant 42471414, Grant 42071412, and Grant 424B2009; in part by the Fundamental Research Funds for the Central Universities under Grant 2042024kf0020 and Grant 2042023kfyq04; in part by the Natural Science Foundation of Heilongjiang Province of China for Excellent Youth Project under Grant YQ2024F006; and in part by the Open Research Fund from Guangdong Laboratory of Artificial Intelligence and Digital Economy (SZ) under Grant GML-KF-24-09. The associate editor coordinating the review of this article and approving it for publication was Dr. Siwei Ma. *(Corresponding authors: Qiangqiang Yuan; Kui Jiang.)*

Yi Xiao, Qiangqiang Yuan, and Yuzeng Chen are with the School of Geodesy and Geomatics, Wuhan University, Wuhan 430072, China (e-mail: xiao_yi@whu.edu.cn; yqiang86@gmail.com; yuzeng_chen@whu.edu.cn).

Kui Jiang is with the School of Computer Science and Technology, Harbin Institute of Technology, Harbin 150000, China (e-mail: jiangkui@hit.edu.cn).

Shiqi Wang is with the Department of Computer Science, City University of Hong Kong, Hong Kong, China (e-mail: shiqwang@cityu.edu.hk).

Chia-Wen Lin is with the Department of Electrical Engineering and the Institute of Communications Engineering, National Tsing Hua University, Hsinchu 30013, Taiwan (e-mail: cwlin@ee.nthu.edu.tw).

Digital Object Identifier 10.1109/TIP.2025.3547298

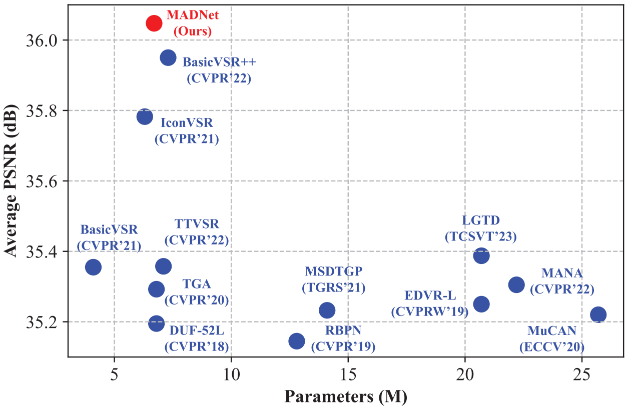


Fig. 1. **Parameters (M) and PSNR (dB) performance comparison.** Our MADNet outperforms state-of-the-art VSR methods with a favorable model size. Comparisons are conducted on the JiLin-1 test set.

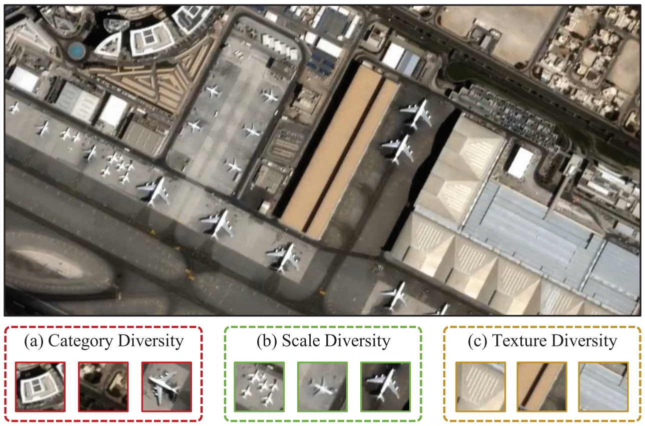


Fig. 2. **Three typical diversities of observed objects:** (a) category diversity (*e.g.*, urban buildings, forests, airplanes), (b) scale diversity (*e.g.*, varying sizes of airplanes), and (c) texture diversity (*e.g.*, smooth surface vs. irregular stripe patterns). Satellite imagery is from Carbonite-2.

observed objects, which require heterogeneous representation for accurate reconstruction, but are barely explored. The reason lies in that the existing VSR methods encode spatial information via static and spatially invariant convolutions, resulting in limited capability to represent various spatial patterns in remote sensing scenarios. **Secondly, the pooling operation in attention operation inevitably causes high-frequency information loss during feature propagation.** The efforts for inter-channel relationship learning could facilitate the global feature fusion, but the pooling operation learns the scalar representation of a channel, making it hard to well capture complex information for various inputs. Therefore, a natural question arises: *whether a more elaborate framework be developed to encode the diverse spatial patterns while preserving the high-frequency textures for accurate spatial-temporal aggregation?*

To tackle these limitations, we introduce the diverse representation of spatial features and harmonize it with the frequency-based multi-spectral channel fusion to explore spatial-temporal information. Specifically, we propose a novel multi-axis feature diversity enhancement module (MADM). Its key component involves three branches to synergistically encode spatial-temporal information: 1) a spatial diversity enhancement (SDE) branches with a set of learnable filters to explore various local patterns; 2) a channel diversity enhancement (CDE) module that replaces the pooling operation with the discrete cosine transform (DCT) to boost the frequency representation in the channel dimension; 3) an auxiliary branch (Aux) that exploits lightweight static convolution to extract spatially invariant features. Unlike previous works that employ dynamic convolution to encode spatial features, we integrate the diverse spatial representation for high-quality feature embedding, followed by the learnable fusion coefficients to achieve adaptive spatial-temporal aggregation. In this manner, the diverse feature patterns are effectively fused with the corresponding weights. By parallel incorporating the proposed SDE, CDE, and Aux into MADM, the spatial variant, frequency diversity, and invariant patterns can be fully explored to produce an enhanced heterogeneous feature representation. These strategies form a novel and effective remote sensing VSR framework, termed MADNet, which achieves state-of-the-art performance on four mainstream satellite video benchmarks.

To sum up, the contributions of this paper are as follows:

1) We propose a novel multi-axis feature diversity enhancement network (MADNet), which explores the diverse spatial and channel feature representation for high-quality remote sensing VSR tasks.
2) We construct the SDE and CDE to boost the feature representation. The former explores the spatially-variant information in satellite videos via a series of learnable kernel bases and linearly fusing design. The latter learns the inter-frequency correlations in the frequency domain with DCT, which helps to preserve more high-frequency cues during long-range propagation.
3) Extensive experiments on four video satellites demonstrate the favorable VSR performance of MADNet, surpassing the SOTA BasicVSR++ by 0.24dB on the JiLin-1 dataset, as shown in Fig. 1.

The remainder of this paper is organized as follows: Section II reviews the progress of natural and remote sensing VSR, as well as some related works to this study. Section III presents implementation details of the proposed MADNet. Section IV contains rigorous experiments and ablation analysis, and we summarize the whole paper in section V.

II. Related Work

A. Natural Video Super-Resolution

The key to VSR lies in how to explore spatial-temporal information. Broadly, existing VSR methods can be divided into two categories: sliding-window and recurrent methods. Here, we initially review sliding-window VSR approaches that explore accessible information within a local window. Then, we introduce recurrent VSR models, where global redundancy can propagate sequentially.

1) Sliding-Window Methods: Most early studies utilized window sliding to aggregate information from multiple neighboring frames to a target frame. They generally follow the following paradigm: alignment, fusion, and reconstruction. Depending on the alignment process, these studies can be further categorized into explicit and implicit alignment methods. The former usually utilizes optical flow warping to realize frame or feature-wise compensation. To generate accurate optical flows, various methods have been adopted, such as numerical solutions [29], [30], pre-trained models [31], [32], [33], and learnable networks [34], [35]. Huang et al. [31] compensated the target images with a total variation-based optical flow strategy. Ilg et al. [32] introduced the Druleas algorithm to estimate the flow maps. However, they face harsh iterative solving processes and are time-consuming. Wang et al. [35] developed a learnable sub-network to predict the latent high-resolution optical flows, thus providing more HR cues for compensation. Shi et al. [36] directly computed optical flows from the pre-trained SPyNet [33], which offers decent performance and efficiency.

Implicit alignment methods typically incorporate the alignment process into adaptive convolution kernels, with remarkable success in 3D convolution [22], [37], deformable convolution [18], [38], and non-local convolution [23], [39]. Jo et al. [22] proposed directly learning 3D upsampling filters to upscale LR frames to the HR space. Wen et al. [37] proposed a spatio-temporal 3D convolution to realize adaptive alignment. Tian et al. [38] predicted offset parameters to guide the deformable convolution process. Wang et al. [18] put forward a pyramid structure to generate accurate offsets. Some works integrated optical flow into deformable convolution to tame the training difficulties. Yi et al. [39] employed non-local convolution to aggregate global redundancy. Recently, Mei et al. [23] further designed an improved multi-memory non-local attention to cash more useful information.

In summary, despite achieving impressive success, these methods neglect the potential information available in long-distance frame sequences, thus reaching a performance plateau.

2) Recurrent Methods: Due to its recurrent nature, the recurrent approach allows for sequential information propagation, mitigating the aforementioned issues comfortably. Based on the propagation direction, they can be grouped into two types: unidirectional and bidirectional approaches. The former often propagates spatial-temporal information in a single direction. Sajjadi et al. [25] proposed to exploit the previous super-resolved HR frame to reconstruct the next LR frame. Isobe et al. [20] proposed a detailed persevering block to alleviate the error accumulation in recurrent structure. However, single-direction propagation leads to imbalanced information utilization, as early frames naturally have insufficient available information.

The bidirectional propagation model propagates the forward and backward information independently, making the propagation mode effective. Chen et al. [19] investigated the generic framework of VSR and proposed a BasicVSR model that adopts bidirectional propagation. By incorporating the information-refill strategy, they further put forward IconVSR. Later, they developed a second-order propagation strategy and proposed BasicVSR++. Recently, Liu et al. [28] proposed trajectory-aware attention, which grasps global dependence via the self-attention mechanism.

To sum up, the recurrent approach can fully utilize long-distance information and achieve state-of-the-art (SOTA) fashion. Nevertheless, they all use static convolution (*e.g.,* residual block) to encode spatial-temporal information during propagation, lacking inadequate consideration of feature diversity in remote sensing scenarios. This weakness results in suboptimal feature representation, making spatial-temporal aggregation less effective.

B. Remote Sensing Video Super-Resolution

Currently, VSR for satellite video is in its infancy. Early works [40], [41], [42] usually super-resolve satellite video with SISR models, lacking enough exploration in the spatial-temporal dimension. Subsequently, some VSR methods began to appear. Typically, more efforts have been paid to consider the unique characteristics of remote sensing imagery to enhance the feature representation capability. Liu et al. [43] choose to realize alignment using patch similarity as pixel-wise redundancy is hard to extract in satellite videos. Later, they developed a practical scheme for satellite VSR under multiple degradations [44]. He and He [45] equipped the 3D convolution with multi-scale design, enhancing the multi-scale diversity. Similarly, Xiao et al. [46] proposed a multi-scale deformable convolution alignment module to explore the multi-scale redundancy in satellite videos. They also proposed to achieve efficient alignment between satellite video frames via the straightforward temporal difference [24]. Recently, Ni and Zhang [47] adopted deformable convolution and proposed a continuous-scale VSR network.

Nevertheless, all these methods adopt the window-sliding fashion, which is insufficient in capturing valuable long-distance dependency. In this paper, we proposed a recurrent framework for satellite VSR and focused on enhancing the feature diversity to boost the spatial-temporal aggregation performance for better reconstruction.

C. Dynamic Networks

To increase the feature diversity, another line of research attempted to develop dynamic networks to aggregate features adaptively. Bako et al. [48] proposed a kernel prediction network to make filtering kernel more complicated and general by estimating the local weighting kernels. Mildenhall et al. [49] introduced the idea of predicting spatially varying kernels in denoising tasks. Recently, Jiang et al. [50] exploited a smaller set of spatially-varying convolution kernels generated from an efficient predictor to reach a compromise between static and dynamic, as kernel prediction is of high model complexity.

Despite they can enhance the feature diversity, they often use complex network design to estimate the dynamic convolutional filters, which inevitably increase the computational budgets. Moreover, dynamic networks face harsh optimization problems, making them less efficient in VSR tasks. In contrast,

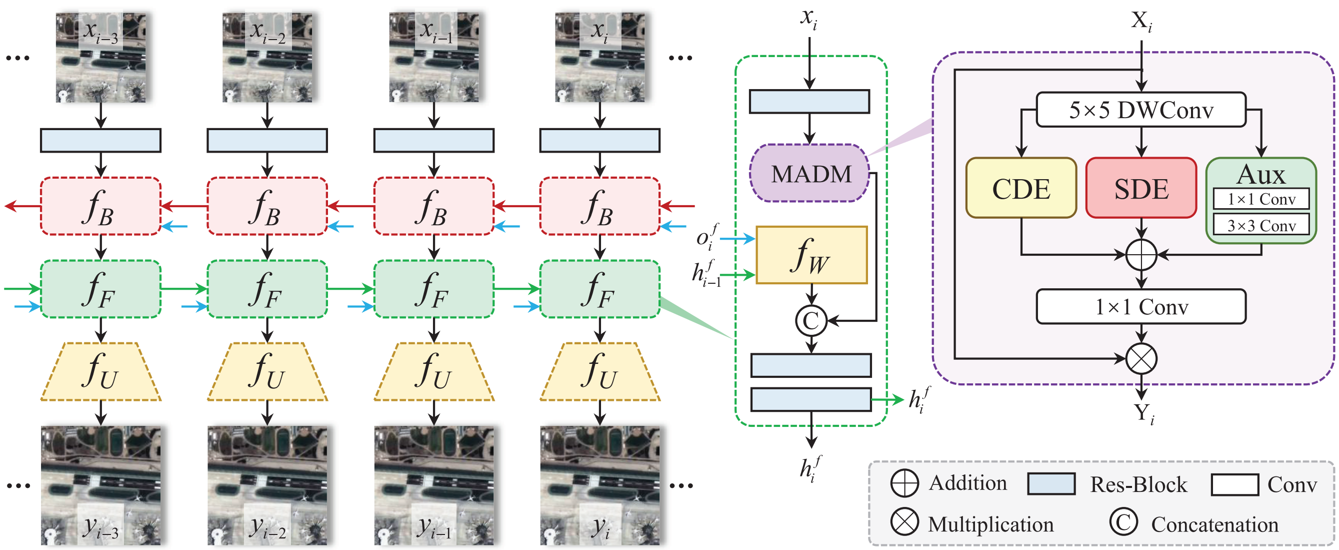


Fig. 3. **An overall of the proposed MADNet.** The red and green arrows denote backward and forward propagation, respectively. The blue arrows mean optical flows. MADNet adopts the generic components of BasicVSR, where f_B, f_F, f_U, and f_W denote the forward and backward propagation modules, the upsampling module, and feature-wise flow warping, respectively. The purple block denotes the proposed multi-axis feature diversity enhancement module (MADM). Channel Diversity Enhancement (CDE), Spatial Diversity Enhancement (SDE), and an auxiliary (Aux) branch process the input feature parallelly. The outputs of these branches are merged and used as an activation to rescale the input feature. DWConv means depth-wise convolution.

we aggregate a set of learnable kernels with linear fusion, which maintains the merits of spatially invariant exploration while greatly reducing model consumption. In addition to spatial diversity, we also perform diversity in the frequency domain, achieving multi-axis diversity enhancement.

III. Proposed Method

In this section, we first present an overview of the proposed MADNet. We then introduce the key components of the multi-axis diversity enhancement module (MADM), *i.e.,* spatial diversity enhancement (SDE) that learns spatially-variant patterns, and channel diversity enhancement (CDE) that explores diverse inter-frequency correlations. Finally, we elaborate on the design pipeline of MADM based on the SDE and CDE.

A. Overview

The overall flowchart of our proposed MADNet is illustrated in Fig. 3. Given an LR satellite video input, residual blocks are adopted for feature extraction. Although residual blocks neglect the feature diversity, they can encode the vital invariant representation. We adhere to the generic framework of IconVSR [19], employing bidirectional propagation through the backward (f_B) and forward (f_F) propagation modules. Note that MADM is integrated into both f_B and f_F. Within MADM, diversity enhancement takes place across multiple branches, including SDE, CDE, and the Auxiliary branches. Following the multi-axis exploration, the outputs are aggregated and utilized to modulate the shallow feature to enhance the representation of spatio-temporal features. The widely-used flow wrapping (f_W) is applied to align with the previous state h_{i-1}, where the optical flow maps are generated by pre-trained SPyNet [33]. The compensated state is then concatenated with the enriched spatio-temporal feature for the upcoming deep feature extraction. We employ a residual block and a pixel-shuffle layer for the final upsampling (f_U) process.

B. Spatial Diversity Enhancement Module

How to gather enriched spatial-temporal representations holds significant importance in VSR tasks. Most prior works rely on spatially invariant convolutions to encode features, where the kernel weights are fixed throughout the process. However, they fail to adaptively grasp the spatial diversity inherent in remote sensing scenarios. In contrast, dynamic convolution introduces a mechanism where the convolutional kernels adapt based on the input features, enabling the model to learn different filters for different regions of the input. To this end, we propose a spatial diversity enhancement module to characterize representative spatial patterns that employs various learnable filters to capture heterogeneous spatial patterns. We then fuse these spatial patterns adaptively to learn distinct convolution weights for individual pixels, thus facilitating the representation of spatial diversity.

The proposed SDE diverges from existing dynamic convolution approaches [50], [51] or kernel prediction networks [48] by not directly predicting all kernel weights. Instead, the aggregation weights of the SDE module are determined through spatially adaptive fusion of shared kernel bases. Consequently, our SDE module is lighter and more straightforward to optimize. Ablation studies in Table V confirm the effectiveness of the proposed design choices against dynamic convolution.

Specifically, as shown in Fig. 4(a), given a shallow feature $X_{in} \in \mathbb{R}^{h \times w \times c}$, where h, w, and c respectively represent the height, width, and channel numbers, our SDE set N learnable filters to capture diverse spatial patterns. Here, $K \in \mathbb{R}^{N \times c \times k^2}$ comprises N convolution kernels with a kernel size of k. To incorporate spatial diversity extracted by these filters,

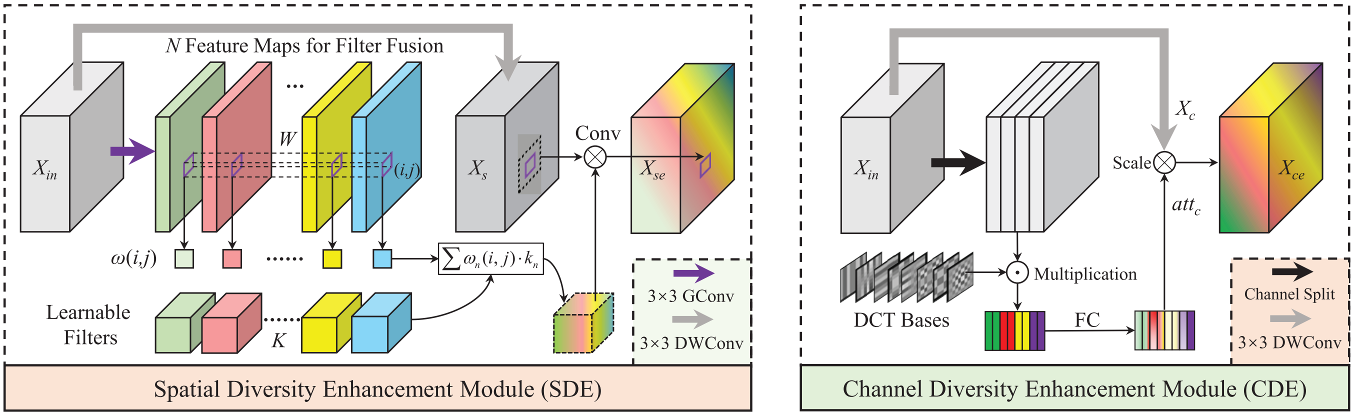


Fig. 4. **The details of the proposed SDE and CDE.** SDE receives an input X_{in}, then SDE predicts N feature maps to fuse diverse filters, generating an enriched convolution weight for a specific location. The fused filters can adaptively grasp the spatial diversity of the encoded input feature to generate a spatially-invariant output X_{se}. On the other hand, with the input X_{in}, CDE starts with diverse DCT bases to generate diverse frequency components. It then utilizes a fully connected layer to produce channel activation, thus adaptively enhancing the channel-wise diversity in the frequency space.

we adaptively predict N feature maps as fusion coefficients $W \in \mathbb{R}^{N\times h\times w}$ using a lightweight 3×3 grouped convolution (GConv). This approach enables the adaptive combination of diversity filters to derive an enriched convolution weight for each spatial location.

In particular, for a spatial location (i, j), the corresponding value of the n-th feature map in the feature map W is used to scale the n-th filter. Mathematically, the enriched weight E can be obtained as follows:

$$E(i, j) = \sum_{n=1}^{N} \omega_n(i, j) \cdot k_n, \tag{1}$$

where k_n represents the n-th learnable kernel, and $\omega_n(i, j)$ denotes the n-th filter fusion coefficients at location (i, j).

To further transform the features, the input feature X_{in} is converted to an embedding X_s by a 5×5 DWConv for adaptive convolution with enriched convolution weight E. After the convolution operation $\otimes$, the spatially enhanced feature X_{se} at location (i, j) can be written as

$$X_{se}(i, j) = X_s^R(i, j) \otimes E(i, j), \tag{2}$$

where $X_s^R(i, j)$ is a $k\times k$ convolution sampling region surrounding location (i, j).

C. Channel Diversity Enhancement Module

Previous works often exploit channel attention (CA) mechanisms [55] to explore the intra-channel relationship. Nevertheless, the global pooling operation used in CA inevitably discards the high-frequency diversity of feature maps, resulting in suboptimal performance in representing the channels. As demonstrated in previous works [56], transforming spatial features into the frequency domain can significantly aid in reconstructing high-quality textures. However, non-parametric frequency transformation methods, such as the fast fourier transform (FFT), are often sensitive to noise and variations in image scale factors [57]. To address these issues and more effectively compress channels while preserving important frequency patterns, we propose utilizing the discrete cosine transform (DCT), which offers a more robust and efficient approach for capturing these frequency components. As illustrated in Fig. 4(b), given an input feature X_{in}, it is split into n chunks in the channel dimension. Denote the i-th part as $X^i \in \mathbb{R}^{H\times W\times d}$, where $d = C/n$.

DCT has been widely used to convert spatial features to the frequency domain. Mathematically, the two-dimensional (2D) DCT can be written as:

$$T_{h,w}^{i,j} = \cos\left(\frac{\pi h}{H}\left(i + \frac{1}{2}\right)\right) \cos\left(\frac{\pi w}{W}\left(i + \frac{1}{2}\right)\right), \tag{3}$$

where H and W are the height and width of input feature X_{in}. For simplicity, we denote d-th 2D DCT base as T_k. By applying T_k to X_{in}, we obtain the transformed frequency spectrum, which means:

$$Freq^k = \sum_{h=0}^{H-1}\sum_{w=0}^{W-1} X_{h,w}^k T_{h,w}^k, \tag{4}$$

where $Freq^k$ is the d dimensional flatten vector after compression.

Ultimately, the multiple vectors yield from diverse frequency components are merged and activated as:

$$att_c = \sigma(FC([Freq^1, Freq^2, \cdots, Freq^n])), \tag{5}$$

where σ is sigmoid function, FC means a fully connected layer and $[\cdot]$ presents channel-wise concatenation.

D. Multi-Axis Diversity Enhancement Module

To comprehensively encode diversity features for satellite VSR, based on the SDE and CDE, we designed a multi-axis diversity enhancement module (MADM) to handle the spatial and channel heterogeneous information. As shown in Fig. 3, MADM initiates with a 5×5 DWConv for feature extraction. Subsequently, SDE, CDE, and an auxiliary branch explore spatial variant, channel diversity, and spatially-invariant features, respectively. In addition to diverse spatial and frequency patterns, spatially-invariant features are also vital for reconstruction. To introduce static spatial-temporal representations

TABLE I

QUANTITATIVE COMPARISONS IN TERMS OF PSNR AND SSIM. EIGHT SCENES OF JILIN-1 ARE SELECTED AND THE RESULTS ARE CALCULATED ON THE LUMINANCE CHANNEL (Y). THE BEST AND SECOND PERFORMANCES ARE HIGHLIGHTED IN RED AND BLUE, RESPECTIVELY

Methods	Scene-1	Scene-2	Scene-3	Scene-4	Scene-5	Scene-6	Scene-7	Scene-8
Bicubic	31.05/0.9097	29.69/0.8795	31.97/0.9242	32.28/0.9251	28.57/0.8630	30.62/0.9009	33.72/0.9391	30.83/0.8989
DUF-52L [22]	35.83/0.9604	33.54/0.9450	36.72/0.9646	36.90/0.9664	32.28/0.9326	34.57/0.9525	39.15/0.9761	34.60/0.9489
RBPN [30]	35.73/0.9595	33.46/0.9440	36.56/0.9637	36.62/0.9647	32.10/0.9307	34.65/0.9534	39.26/0.9767	34.61/0.9487
SOF-VSR [35]	35.89/0.9610	33.53/0.9453	36.75/0.9643	36.86/0.9656	32.26/0.9328	34.70/0.9537	39.15/0.9760	34.74/0.9498
EDVR-L [18]	36.12/0.9624	33.67/0.9469	37.00/0.9661	37.08/0.9675	32.43/0.9349	34.89/0.9555	39.42/0.9772	34.99/0.9526
MuCAN [52]	35.96/0.9612	33.40/0.9448	36.88/0.9649	36.73/0.9656	32.20/0.9330	34.68/0.9543	39.30/0.9768	34.61/0.9495
TGA [53]	36.07/0.9636	33.68/0.9484	37.04/0.9679	36.98/0.9678	32.38/0.9370	34.86/0.9563	39.11/0.9765	34.93/0.9531
MSDTGP [46]	36.13/0.9631	33.55/0.9462	36.74/0.9646	36.92/0.9674	32.42/0.9350	34.81/0.9551	39.46/0.9774	34.76/0.9512
MANA [23]	35.94/0.9616	33.71/0.9470	36.78/0.9641	36.89/0.9662	32.43/0.9347	34.78/0.9544	39.34/0.9768	34.81/0.9509
LGTD [24]	36.21/0.9626	33.80/0.9484	36.95/0.9655	36.97/0.9671	32.58/0.9369	34.93/0.9559	39.47/0.9773	34.89/0.9522
RRN [26]	35.74/0.9611	33.36/0.9450	36.66/0.9650	36.71/0.9660	32.05/0.9332	34.67/0.9544	39.08/0.9759	34.67/0.9506
BasicVSR [19]	36.14/0.9631	33.88/0.9492	37.08/0.9669	37.17/0.9686	32.60/0.9372	34.87/0.9551	39.40/0.9770	34.91/0.9519
IconVSR [19]	36.60/0.9657	33.82/0.9502	37.43/0.9696	37.43/0.9704	32.77/0.9403	34.91/0.9578	39.56/0.9781	34.85/0.9534
TTVSR† [28]	36.18/0.9637	33.80/0.9486	37.14/0.9677	37.12/0.9679	32.47/0.9363	34.74/0.9548	39.46/0.9773	34.84/0.9522
BasicVSR++ [27]	**36.65/0.9663**	**33.95/0.9503**	**37.67/0.9701**	**37.29/0.9692**	**32.84/0.9403**	**35.35/0.9600**	**39.83/0.9789**	**35.32/0.9563**
RVRT* [54]	36.30/0.9642	33.93/0.9492	37.33/0.9679	37.26/0.9685	32.74/0.9383	35.12/0.9575	39.65/0.9781	35.14/0.9537
MADNet (Ours)	**36.84/0.9674**	**34.29/0.9536**	**37.82/0.9712**	**37.52/0.9707**	**33.07/0.9432**	**35.55/0.9619**	**40.14/0.9801**	**35.65/0.9590**

TABLE II

QUANTITATIVE COMPARISON IN TERMS OF PSNR AND SSIM. THE FLOPS ARE COMPUTED ON AN INPUT TENSOR WITH THE SIZE OF 1×N×3×160×160, WHERE N MEANS THE NUMBER OF INPUT FRAMES. NOTE THAT TTVSR† AND RVRT* DENOTE WE RETRAIN TTVSR AND RVRT BY SETTING THE PROPAGATE LENGTH TO 15 AND 16, RESPECTIVELY. TO SAVE GPU MEMORY MAINTENANCE, WE SET THE EMBEDDING DIMENSION TO 96 IN RVRT*. THE BEST AND SECOND PERFORMANCES ARE HIGHLIGHTED IN RED AND BLUE, RESPECTIVELY

Methods	Frames	#Param. (M)	FLOPs (G)	Publications	JiLin-1	Carbonite-2	UrtheCast	SkySat-1
Bicubic	-	-	-	-	31.09/0.9051	36.23/0.9279	30.48/0.8346	30.74/0.8700
DUF-52L [22]	7	6.8	736.6	CVPR'18	35.45/0.9558	39.19/0.9534	32.46/0.8852	33.68/0.9263
RBPN [30]	5	12.8	3785.3	CVPR'19	35.37/0.9552	39.24/0.9540	32.32/0.8834	33.65/0.9256
SOFVSR [35]	3	1.1	127.3	TIP'20	35.47/0.9561	38.98/0.9519	32.22/0.8821	33.32/0.9221
EDVR-L [18]	5	20.7	897.8	CVPRW'19	35.70/0.9579	39.37/0.9547	32.58/0.8878	33.90/0.9295
MuCAN [52]	5	25.7	3462.0	ECCV'20	35.47/0.9563	39.24/0.9545	32.40/0.8843	33.77/0.9285
MSDTGP [46]	5	14.1	1579.8	TGRS'21	35.60/0.9575	39.22/0.9537	32.41/0.8858	33.70/0.9272
TGA [53]	7	7.1	-	CVPR'20	35.63/0.9588	39.40/0.9547	32.43/0.8870	33.97/0.9312
MANA [23]	7	22.2	633.5	CVPR'22	35.59/0.9570	39.32/0.9541	32.49/0.8869	33.82/0.9284
LGTD [24]	5	20.7	647.8	TCSVT'23	35.73/0.9582	39.23/0.9546	32.31/0.8840	33.73/0.9286
RRN [26]	7	3.4	-	BMVC'20	35.37/0.9564	39.10/0.9538	32.15/0.8836	33.52/0.9266
BasicVSR [19]	15	4.1	1631.6	CVPR'21	35.75/0.9586	39.40/0.9549	32.38/0.8858	33.89/0.9291
IconVSR [19]	15	6.3	2266.3	CVPR'21	35.90/0.9607	40.03/0.9608	32.48/0.8881	34.72/0.9409
TTVSR†[28]	15	6.8	4429.9	CVPR'22	35.72/0.9586	39.29/0.9549	32.35/0.8848	33.81/0.9290
BasicVSR++ [27]	30	7.3	2719.4	CVPR'22	36.11/0.9614	**40.11/0.9610**	**32.59/0.8884**	**34.99/0.9432**
RVRT* [54]	16	6.2	2643.9	NIPS'22	35.93/0.9597	39.57/0.9567	**32.63/0.8887**	33.99/0.9311
IA-RT [58]	16	13.4	-	CVPR'24	**36.17/0.9618**	40.07/0.9607	32.57/0.8885	34.74/0.9413
MADNet (Ours)	15	6.7	2444.5	-	**36.35/0.9634**	**40.21/0.9612**	32.51/0.8881	**35.11/0.9449**

into MADM, we integrate a 1 × 1 convolution followed by a 3 × 3 convolution in the auxiliary branch. This lightweight design enables MADM to explore both static and dynamic dependencies, thus enhancing the spatial-temporal aggregation for accurate VSR. Intuitively, addition and multiplication are commonly used operations for feature aggregation. However, the diversity from each branch still exists a pattern gap.

Instead of merely aggregating the output of these branches and supplementing it to X_{in} with a global connection, we opt for a progressive aggregation approach in two stages. Later, the second stage employs point-wise multiplication to modulate the input feature. This two-stage aggregation introduces additional non-linear activation in the MADM module.

IV. EXPERIMENT AND DISCUSSION

A. Satellite Video Datasets and Evaluation Metrics

1) Satellite Video Datasets: To comprehensively evaluate the VSR performance on remote sensing scenarios, we collected extensive satellite videos from four mainstream video satellites, including JiLin-1, Carbonite-2, SkySat-1, and UrtheCast. The original frame size varies from 1920 × 1080 to 4096 × 2160, and the duration is from 12s to 60s. Following previous work [46], we extract 189 satellite video clips from JiLin-1 to build our training set, termed JiLin-189, where each sub-clip contains 100 consecutive frames with a resolution of 640 × 640. Besides model training, eight scenes are randomly cropped from JiLin-1 satellite videos for model testing. Note that the training and test clips are captured from different ground regions and do not overlap with each other.

Additionally, we further randomly crop 10, 6, and 12 scenes from Carbonite-2, SkySat-1, and UrtheCast, respectively, to form another three test sets, where each video includes 100 frames of image size 512 × 512. To sum up, a total of 36 scenes from four video satellites are used to evaluate VSR performance. The original video clips are downsized to 1/4 of

TABLE III

QUANTITATIVE RESULTS IN TERMS OF LPIPS [62]. THE BEST AND SECOND PERFORMANCES ARE HIGHLIGHTED IN RED AND BLUE, RESPECTIVELY

Datasets	Bicubic	EDVR-L [18]	MANA [23]	RRN [26]	BasicVSR [19]	IconVSR [19]	TTVSR† [28]	BasicVSR++ [27]	RVRT* [54]	MADNet
JiLin-1	0.2748	0.1165	0.1136	0.1164	0.1183	**0.0941**	0.1071	0.0993	0.1116	**0.0904**
Carbonite-2	0.3010	0.1791	0.1773	0.1715	0.1728	0.1550	0.1715	**0.1492**	0.1715	**0.1433**
SkySat-1	0.3240	0.1395	0.1399	0.1455	0.1397	0.1202	0.1297	**0.1162**	0.1371	**0.1085**
UrtheCast	0.3168	0.1853	0.1836	0.1851	0.1829	0.1754	0.1784	**0.1719**	0.1831	**0.1709**
Average	0.3041	0.1551	0.1536	0.1546	0.1534	0.1362	0.1467	**0.1341**	0.1508	**0.1283**

TABLE IV

QUANTITATIVE COMPARISON ON REDS4 [63] AND VID4 [64] IN TERMS OF PSNR AND SSIM. THE BEST AND SECOND-BAST PERFORMANCES ARE HIGHLIGHTED IN BOLD AND UNDERLINED, RESPECTIVELY

Methods	Propagation	REDS4 [63]		Vid4 [64]	
		PSNR	SSIM	PSNR	SSIM
DUF-52L [22]	First Order	28.63	0.8251	27.33	0.8319
RBPN [30]		30.09	0.8590	27.12	0.8180
EDVR-L [18]		31.09	0.8800	27.35	0.8264
MuCAN [52]		30.88	0.8750	-	-
BasicVSR [19]		31.42	0.8909	27.24	0.8251
IconVSR [19]		31.67	0.8948	27.39	0.8279
MADNet (Ours)		**32.43**	**0.9084**	**27.86**	**0.8421**
BasicVSR++ [27]	Second Order	32.39	0.9069	27.79	0.8400
VRT [65]		32.19	0.9006	27.93	0.8425
RVRT [54]		**32.75**	**0.9113**	**27.99**	**0.8462**

the original size using *bicubic* interpolation, generating paired LR-HR satellite videos for ×4 VSR.

2) Evaluation Metrics: Similar to prior works [59], [60], the widely used Peak Signal-to-Noise Ratio (PSNR) and Structural Similarity index (SSIM) [61] are adopted to measure the fidelity between reconstructed and ground-truth videos. Besides, the Learned Perceptual Image Patch Similarity (LPIPS) [62] is used to evaluate the perceptual quality of restored videos. The lower LPIPS values indicate better perceptual performance. Notably, the PSNR and SSIM are calculated on the Y channel of the YCbCr image space.

B. Implementation Details

1) Model Details: MADNet only implements ×4 VSR. The inter-channel number of our MADNet is set to 64. Note that we adopt the pre-trained SPyNet [33] to estimate the optical flow maps. In addition to SPyNet, our MADNet does not involve any pre-training process. Each forward (f_F) and backward module (f_B) contains 7 residual blocks. In the SDE, the number of learnable filters is set to 128, and the kernel size is 3×3. As for CDE, we select four DCT bases for channel compression, which means we split the feature into 4 groups in the channel dimension. The most important 32 channels are maintained in CDE. The channel number of the two fully connected layers is 256 and 64, respectively.

2) Training Details: We adopt the Adam optimizer with $\beta_1 = 0.9$ and $\beta_2 = 0.99$ for model training. The batch size is set to 4. The learning rate is initialized to 2×10^{-4} and updated by the cosine annealing schedule. Following previous works, we employ 15 video frames for bidirectional propagation. The input size of the LR videos is 64×64. During model training, data augmentation is performed by random rotation (90Â°, 180Â°, 270Â°) and flipping (vertical or horizontal). MADNet is trained with a total of 100,000 iterations to reach robust convergence. The widely used Carbonnier penalty loss is adopted as the training object. In particular, $\mathcal{L} = \sqrt{\|x_t - y_t\|^2 + \varepsilon^2}$, where $\varepsilon = 10^{-3}$. Our MADNet is implemented with the PyTorch framework and trained on a single NVIDIA RTX 3090 GPU with 24GB memory and a 3.40 GHz AMD Ryzen 5700X CPU. It takes nearly three days to train our MADNet.

C. Comparison With State-of-the-Art Methods

To evaluate the performance of the proposed MADNet, we compared MADNet with representative approaches, including Bicubic Interpolation, sliding-window VSR methods (DUF-52L [22], RBPN [30], SOF-VSR [35], EDVR-L [18], MuCAN [52], MSDTGP [46], TGA [53], MANA [23], LGTD [24]), and recurrent VSR models (RRN [26], BasicVSR [19], IconVSR [19], TTVSR [28], BasicVSR++ [27], RVRT [54], and IA-RT [58]). For a fair comparison, we retrained these methods from scratch on the JiLin-189 training set following their official implementation settings. Notably, limited by the CUDA memory, the frame length used for propagation of TTVSR and RVRT is set to 15 and 16, respectively, and we denote them TTVSR† and RVRT*. Following the test setting in DUF-52L [22], 8 pixels on the boundary are cropped before metric calculation.

1) Quantitative Comparison: The PSNR/SSIM results for eight scenes in the JiLin-1 test set are listed in Table. I. From Table I, we can see that our MADNet outperforms the SOTA sliding-window and recurrent VSR methods across all eight scenes. For instance, MADNet exhibits a substantial lead over BasicVSR++ in Scene-2 (24.29dB vs. 33.95dB), indicating its superior VSR performance on satellite videos.

Moreover, the average PSNR/SSIM results on JiLin-1, Carbonite-2, SkySat-1, and UrtheCast are reported in Table II, along with input frame number, parameters, and FLOPs. From this table, it is observed that MADNet consistently demonstrates favorable performance across various video satellite platforms, which demonstrates the generalization and robustness of MADNet. In particular, MADNet outperforms EDVR-L, a large-capacity sliding-window approach, by up to 0.84dB in PSNR on the Carbonite-2 test set while employing 68% fewer parameters. In the SkySat-1 test set, MADNet surpasses the impressive recurrent approach BasicVSR++ in PSNR (35.11dB vs. 34.99dB) with fewer parameters (6.7M vs. 7.3M) and FLOPs (2444.5G vs. 2719.4G). When compared to the SOTA method IconVSR, MADNet demonstrates high performance with gains of 0.45dB, 0.18dB, 0.03dB, and 0.39dB on JiLin-1, Carbonite-2, UrtheCast, and SkySat-1 test sets, respectively. Due to the limited generalization capability

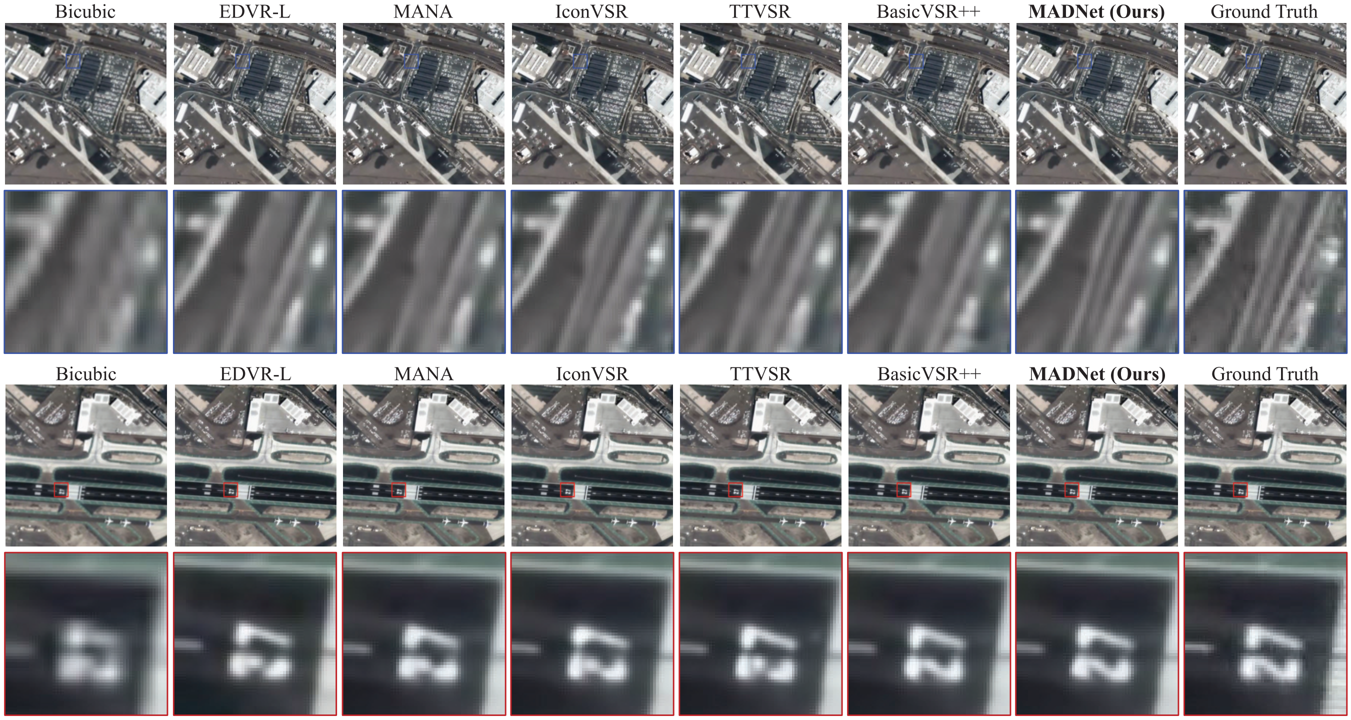


Fig. 5. **Qualitative comparisons.** Scene-3 (top) and Scene-2 (bottom) are selected from JiLin-1 test set. Zoom in for better visualization.

and the domain gap between Jilin-1 and Urthecast, MADNet's performance is less pronounced compared to BasicVSR++ on UrtheCast. These results demonstrate the favorable performance and generalization capability of MADNet across various video satellite platforms. Note that both IconVSR and BasicVSR++ employ static convolution for aggregating spatial-temporal information, overlooking the feature diversity present in satellite videos. Our MADNet benefits from a more enriched spatial-temporal feature by SDE and CDE, leading to superior VSR performance.

The LPIPS results are reported in Table. III. From Table III, it can be observed that our MADNet achieves the best performance on all test sets. For instance, MADNet achieves 0.1283 in LPIPS on average, which is better than EDVR-L and BasicVSR++ by 0.0268 and 0.0058, respectively. These observations highlight the favorable capability of our MADNet in recovering videos with high perceptual quality.

Moreover, we further retrained MADNet on REDS [63] to evaluate its performance on natural video benchmarks [64]. Table IV presents a quantitative comparison with SOTA VSR methods. Specifically, for first-order propagation models, MADNet achieves the best performance. It outperforms IconVSR by 0.76dB on PSNR and 0.0136 on SSIM for REDS4, demonstrating its effectiveness in natural VSR tasks. This is because recovering high-frequency information is also crucial for natural VSR. Benefiting the global frequency exploration of CDE, our MADNet could restore more sharp cues of natural videos, thus achieving favorable reconstruction results. Compared to the second-order propagation approaches BasicVSR++, our MADNet exhibits superior accuracy in the configuration of first-order propagation. Yet MADNet is slightly below RVRT on Vid4 (27.86dB vs. 27.99dB), primarily due to challenges in information exchange, limiting the benefits of accurate propagation.

2) Qualitative Comparison: In Fig. 5, Fig. 6, and Fig. 7, we provide visual comparisons on various satellite videos to evaluate the qualitative performance of our MADNet. Specifically, as shown in Scene-3 in Fig. 5, MADNet exhibits superior visual quality in recovering details of lines on the ground. BasicVSR++, relying on static convolution for feature encoding, tends to lose high-frequency information during long-distance propagation. In contrast, MADNet explores the diversity in the frequency space, providing more high-frequency cues for better reconstruction. Similarly, in Scene-2 of JiLin-1, MADNet produces clearer and sharper details of the number '27' on the ground compared to other VSR methods. This suggests that both sliding-windows methods and recurrent models are inadequate in providing fine-grained spatial information, particularly in capturing spatial diversity in satellite videos. The incorporation of MADM in MADNet introduces additional heterogeneous features into the VSR model, helping to recover diverse textures in remote sensing scenarios.

As shown in Fig. 6, in Scene-2 of SkySat-1, only our MADNet can recover the correct shape and boundary of objects with spatial diversity. In this context, both EDVR-L and BasicVSR++ produce blurry and inaccurate results. Moreover, as shown in Scene-6 of SkySat-1, MADNet successfully preserves the reliable distribution of edges on the building, while other approaches fail to recover the high-frequency information. In Fig. 7, MADNet consistently produces visually pleasing results compared to other SOTA methods, offering rich spatial details and high-frequency information.

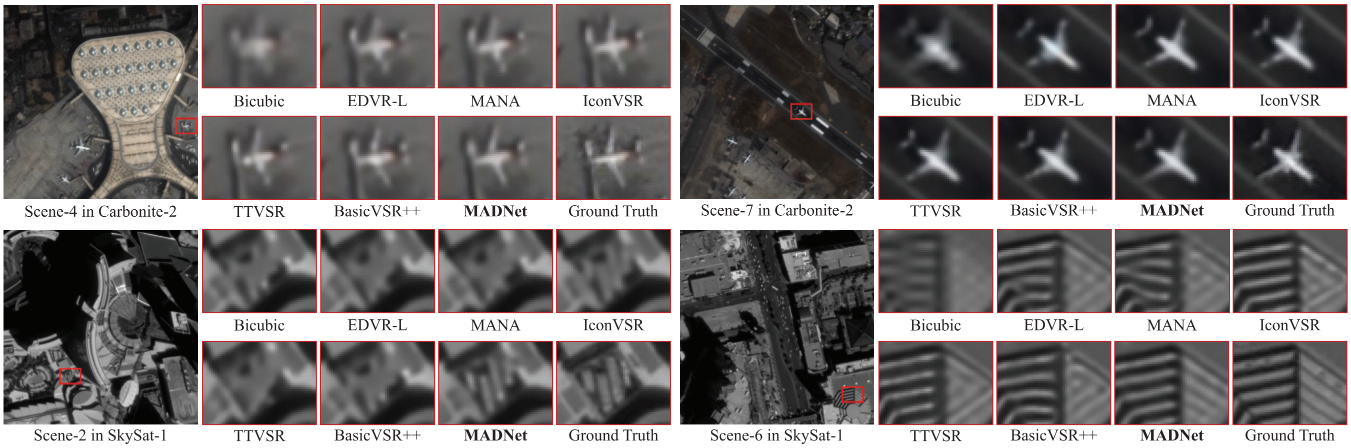


Fig. 6. **Qualitative comparisons.** Scene-4 and Scene-7 are from Carbonite-2 test set while Scene-2 and Scene-6 are from SkySat-1 test set. Zoom in for better visualization.

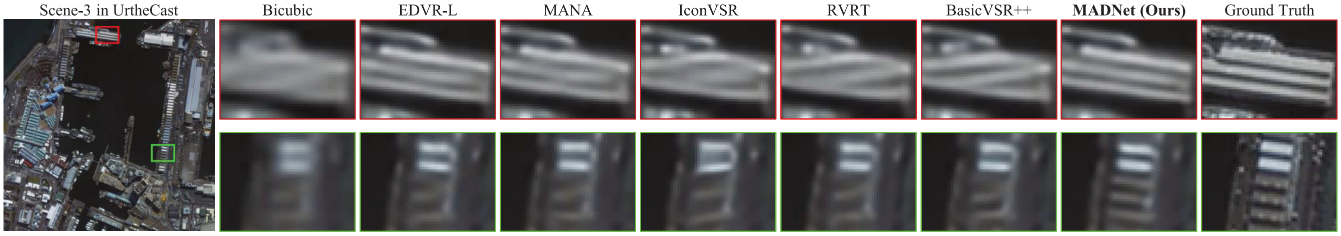


Fig. 7. **Qualitative comparison.** Scene-3 is from the UrtheCast test set. Zoom in for better visualization.

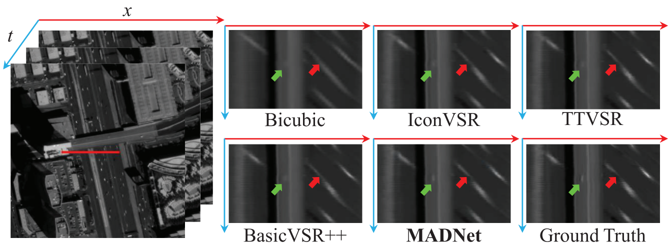


Fig. 8. **Qualitative comparison of temporal profile.** We record a row (red line) and stack it across the timeline to observe the temporal changes. The profile from BasicVSR++ produces a mixed motion pattern (red arrow), while our MADNet exhibits distinctive motions and is close to ground truth. By enhancing the feature diversity, the profile from MADNet exhibits clear details (green arrow).

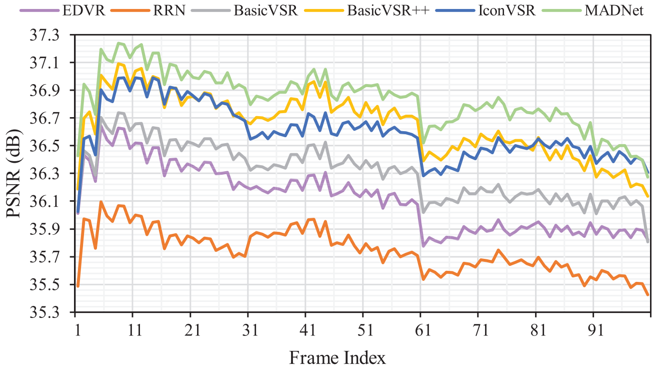


Fig. 9. **Quantitative comparison across frame index.** The results are conducted on Scene-1 of JiLin-1 test set. Our MADNet achieves consistently superior performance on the entire frame.

These visual results demonstrate the effectiveness of SDE in enhancing the representation of different types of structures and textures. Moreover, CDE indeed preserves the vital high-frequency details by exploring the diversity in the frequency domain.

To better evaluate the coherence and consistency of the restored video over long-time series dynamics, we recorded a line of pixels and stacked them on the timeline. The temporal profile is shown in Fig. 8. MADNet demonstrates the ability to generate precise long-term temporal distributions, emphasizing the effectiveness of our MADM in accurate spatial-temporal aggregation. We also evaluate the performance across the entire frame index. The results are illustrated in Fig. 9. MADNet consistently achieves the best performance over the entire frame sequence, demonstrating its capability to aggregate spatial-temporal information for accurate reconstruction.

D. Ablation Studies

This section gives extensive ablation experiments to demonstrate the effectiveness of each component of MADNet. Note that the PSNR is calculated on the JiLin-1 test set.

1) Effectiveness of Individual Components: To verify the effectiveness of the main components in our MADNet, we conduct various ablation studies by progressively incorporating CDE, SDE, and the auxiliary branch into MADM and evaluate the PSNR performance. The quantitative results are reported in Table VI. Compared to the Baseline model, CDE and SDE contribute to a PSNR improvement of 0.24dB and 0.27dB, respectively. contribute to a PSNR improvement of 0.24dB and 0.27dB, respectively. The addition of the

TABLE V
ABLATION OF SDE MODULE. THE PSNR IS TESTED ON THE JILIN-1 TEST SET

Variants	PSNR (dB)
Baseline	36.16
Baseline + Dynamic Conv [51]	36.20
Baseline + SDE	36.35
Baseline + SDE w/. 128 Filters + 3 × 3 DWConv	36.30
Baseline + SDE w/. 128 Filters + 5 × 5 DWConv	**36.35**
Baseline + SDE w/. 128 Filters + 7 × 7 DWConv	36.28
Baseline + SDE w/. 128 Filters + 9 × 9 DWConv	36.32

TABLE VI
ABLATION ON INDIVIDUAL COMPONENTS OF MADNET

Baseline	MADM			#Param.	FLOPs	PSNR (dB)
	CDE	SDE	Aux			
✓				6.43M	2432.98G	35.91
✓	✓			6.66M	2436.57G	36.14
✓		✓		6.67M	2440.95G	36.17
✓	✓	✓		6.67M	2440.95G	36.33
✓	✓	✓	✓	6.68M	2444.54G	**36.35**

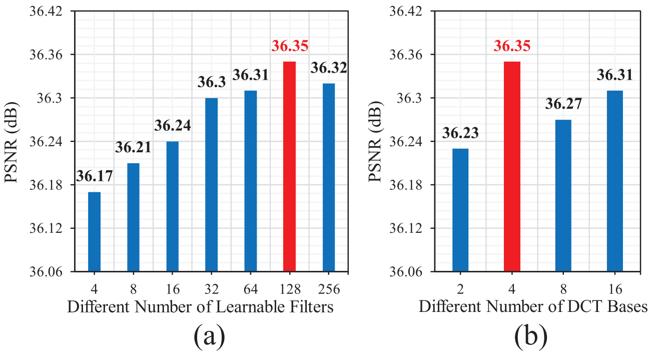


Fig. 10. **Effect of hyperparameters in CDE and SED.** (a) Different number of learnable filters used in SDE. (b) Different number of DCT bases in CDE. The PSNR is tested on JiLin-1 testset.

TABLE VII
ABLATION OF CDE MODULE. THE PSNR IS TESTED ON THE JILIN-1 TEST SET

Variants	PSNR (dB)
Baseline	36.18
Baseline + Channel Attention [62]	36.26
Baseline + CDE	36.35
Baseline + CDE w/. 4 DCT base + 3 × 3 DWConv	36.28
Baseline + CDE w/. 4 DCT base + 5 × 5 DWConv	36.30
Baseline + CDE w/. 4 DCT base + 7 × 7 DWConv	**36.35**
Baseline + CDE w/. 4 DCT base + 9 × 9 DWConv	36.31

TABLE VIII
EFFECT OF DIFFERENT FREQUENCY SELECTION METHODS USED IN CDE. TOP-K MEANS WE RETAIN THE MOST IMPORTANT TOP K% FREQUENCY SIGNALS ALONG THE CHANNEL DIMENSION

Selection Methods	Top-8	Top-16	Top-32	Top-64
PSNR (dB)	36.27	36.30	**36.35**	36.21

auxiliary branch results in a further 0.02dB improvement. As Table VI suggests, the integration of CDE, SDE, and the auxiliary branch yields the best PSNR performance compared to the Baseline (36.35dB vs. 35.90dB), with only a marginal increase in parameters (6.43M vs. 6.68M) and FLOPs (2432.98G vs. 2444.54G).

2) Effectiveness of SDE: To validate the effectiveness of conducting feature diversity enhancement in the spatial dimension, we perform the ablation study of SDE with different variants. The quantitative results are listed in Table V. In particular, we remove the SDE to build a Baseline model. Then, we equip the Baseline with dynamic convolution, revealing a 0.04dB improvement, demonstrating the positive impact of enhancing feature diversity. By comparing the second and third rows in Table V, we can see that SDE outperforms dynamic Conv by 0.15dB in PSNR, indicating that our learnable filter can boost the spatial diversity representation. Besides, we investigate the different number of learnable filters in SDE, and the results are shown in Fig. 10. It can be found that when the number of filters increases, the diversity can be increased progressively and reach a plateau after $N = 32$. We ultimately set $N = 128$ as it delivers the best performance. Finally, we incrementally added different DWConv, such as 3×3, 5×5, 7×7, and 9×9. Experimental results illustrate that DWConv with 5×5 kernel achieves the best performance among other kernels, illustrating a moderate kernel size is necessary for spatial feature embedding.

3) Effectiveness of CDE: To evaluate the effectiveness of channel diversity enhancement, we conduct 7 variants in Table VII. For the study in Table VII, the Baseline model with channel-wise diversity enhancement can achieve favorable performance improvement. In comparing CA and CDE, our CDE outperforms CA by 0.09dB in PSNR, demonstrating the effectiveness of CDE in increasing the frequency diversity for better performance. Inspired by the success of frequency analysis in high-level vision tasks, we introduce frequency channel attention [66] into VSR framework to explore the high-frequency patterns. Considering the weak texture properties of remote sensing videos, we infer that excessive DCT increases the latitude diversity of the channel but simultaneously introduces noise frequency signals. As illustrated in Fig. 10(b), it can be observed that higher DCT numbers do not necessarily lead to better performance in VSR tasks. In contrast to the 16 DCTs used in [66], the proposed CDE achieves optimal performance with only 4 DCTs (36.35 dB vs. 35.31 dB). Ultimately, we chose $N = 4$ in our final CDE. Additionally, we changed the kernel size of DWConv used in CDE. The experimental results indicate that 7×7 kernel size achieves the best PSNR performance, highlighting the importance of a relatively large kernel size for channel feature embedding.

Furthermore, different from the original FacNet, we investigated the impact of different frequency selection methods. The results in Table VIII demonstrate that retaining only 16 channels [66] in the frequency domain tends to result in the loss of critical high-frequency details, resulting in suboptimal

TABLE IX

EFFECT OF DIFFERENT ENHANCEMENT STRATEGIES. THE PSNR RESULTS ARE CALCULATED IN THE JILIN-1 TEST SET

Methods	Stage-1	Stage-2	PSNR (dB)
Model-1	Addition	Addition	36.07
Model-2	Multiplication	Multiplication	36.32
Model-3	Multiplication	Addition	36.24
MADNet (Ours)	Addition	Multiplication	**36.35**

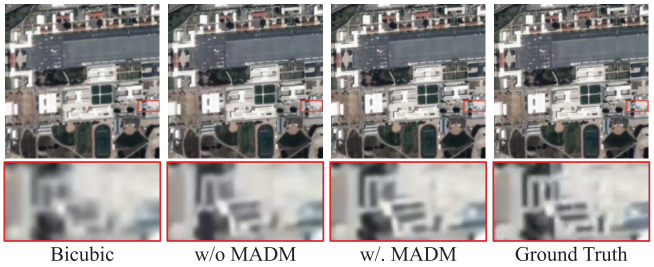


Fig. 11. **Effect of Multi-Axis Diversity Enhancement (f_{MADM}).** The contribution of f_{MADM} is prominent in regions with sharp details, *e.g.,* edge and boundary. The high-frequency information from the additional f_{MADM} leads to favorable improvements.

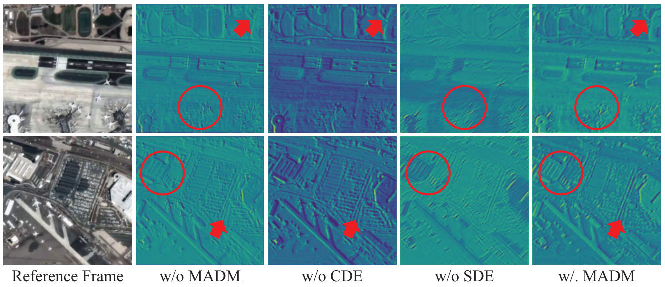


Fig. 12. **Feature visualizations.** Without MADM, the feature experiences blurry details and is not prominent across the entire feature. The feature enhanced by SDE has sharper spatial details, as indicated by the red arrows. The results with CDE activate more high-frequency parts, as indicated by the red circles. The feature enhanced by full MADM is sharp and prominent in fine details.

performance. Meanwhile, selecting all 64 channels introduces undesirable interference signals. Our CDE obtains favorable adaptation, making it applicable to remote sensing VSR tasks.

4) Different Enhancement Operation: To verify the two-stage enhancement operation, we conduct different enhancement operations in stage-1 and stage-2. Empirically, the multiple outputs of Aux, SDE, and CDE can be merged and supplied with addition or activation operations. As listed in Table IX, Model-1 aggregates the outputs and adds them to X_i. Due to the diversity gap between each branch, addition yields the worst PSNR performance. Model-2 adopts point-wise multiplication to fuse the diverse features and modulate X_i with multiplication, achieving a 0.25dB improvement over Model-1. Model-3 uses hybrid operations and is less effective than Model-2. Combining addition and multiplication, our MADNet can further boost PSNR performance by 0.03dB compared to Model-2.

5) Feature Visualization: To better understand how SDE and CDE work, we also visualize the intermediate feature maps of MADNet. As shown in the 2nd and 5th columns of Fig. 12, we observed that the feature map yielded with MADM is cleaner than the feature without MADM. This demonstrates that multi-axis feature diversity enhancement can significantly reduce the noise in features and provide high-quality features for propagation. Besides, as shown in the 3rd column, where we delete CDE and only conduct spatial diversity exploration. We can see that SDE excels at extracting homogeneous spatial patterns (red arrows). Moreover, by comparing the 3rd and 4th columns, we can find CDE succeeds in activating the high-frequency counterparts of the feature map, as highlighted in the red circle. These visualizations suggest that both SDE and CDE effectively enhance feature diversity and provide high-quality representations for improved reconstruction.

V. CONCLUSION

In this paper, we propose MADNet, a novel network developed for satellite VSR. The key design of MADNet is the multi-axis diversity enhancement module, which enhances feature diversity by simultaneously exploring spatially-variant, frequency diversity, and spatial-invariant patterns. To capture various spatial contexts, the proposed SDE module introduces a series of learnable filters to extract different spatial patterns and fuse them adaptively to generate an enriched kernel weight for feature encoding. Additionally, we devise a CDE module, which converts the diversity analysis into the frequency domain with discrete cosine transformation. Extensive experiments on four satellite video benchmarks demonstrate our MADNet achieves favorable performance against SOTA sliding-window and recurrent VSR approaches, demonstrating its superiority both quantitatively and qualitatively.

Despite achieving favorable results under bicubic degradation, the proposed MDANet may collapse in real-world scenes with multiple degradations. Besides, the model complexity of MADNet is still too large for real-time applications. In future work, we plan to extend our MADNet to develop a generalized VSR framework, which offers a broader investigation of real-world satellite videos. Moreover, exploring efficient MADNet would be an exciting direction, providing real-time inference for onboard deployment.

REFERENCES

[1] J. Shao, B. Du, C. Wu, M. Gong, and T. Liu, "HRSiam: High-resolution Siamese network, towards space-borne satellite video tracking," *IEEE Trans. Image Process.*, vol. 30, pp. 3056–3068, 2021.

[2] M. Kaselimi, A. Voulodimos, I. Daskalopoulos, N. Doulamis, and A. Doulamis, "A vision transformer model for convolution-free multilabel classification of satellite imagery in deforestation monitoring," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 34, no. 7, pp. 3299–3307, Jul. 2023.

[3] Y. Yang, X. Tang, Y.-M. Cheung, X. Zhang, and L. Jiao, "SAGN: Semantic-aware graph network for remote sensing scene classification," *IEEE Trans. Image Process.*, vol. 32, pp. 1011–1025, 2023.

[4] H. Guo, Q. Shi, A. Marinoni, B. Du, and L. Zhang, "Deep building footprint update network: A semi-supervised method for updating existing building footprint from bi-temporal remote sensing images," *Remote Sens. Environ.*, vol. 264, Oct. 2021, Art. no. 112589.

[5] J. Li, W. He, W. Cao, L. Zhang, and H. Zhang, "UANet: An uncertainty-aware network for building extraction from remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, 2024, Art. no. 5608513.

[6] Q. He, X. Sun, W. Diao, Z. Yan, F. Yao, and K. Fu, "Multimodal remote sensing image segmentation with intuition-inspired hypergraph modeling," *IEEE Trans. Image Process.*, vol. 32, pp. 1474–1487, 2023.

[7] J. Hou, Q. Cao, R. Ran, C. Liu, J. Li, and L.-J. Deng, "Bidomain modeling paradigm for pansharpening," in *Proc. 31st ACM Int. Conf. Multimedia*, Oct. 2023, pp. 347–357.

[8] Q. Zhang, Q. Yuan, M. Song, H. Yu, and L. Zhang, "Cooperated spectral low-rankness prior and deep spatial prior for HSI unsupervised denoising," *IEEE Trans. Image Process.*, vol. 31, pp. 6356–6368, 2022.

[9] J. Li, K. Zheng, W. Liu, Z. Li, H. Yu, and L. Ni, "Model-guided coarse-to-fine fusion network for unsupervised hyperspectral image super-resolution," *IEEE Geosci. Remote Sens. Lett.*, vol. 20, pp. 1–5, 2023.

[10] J.-N. Su, M. Gan, G.-Y. Chen, W. Guo, and C. L. P. Chen, "High-Similarity-Pass attention for single image super-resolution," *IEEE Trans. Image Process.*, vol. 33, pp. 610–624, 2024.

[11] K. Jiang, Z. Wang, P. Yi, and J. Jiang, "Hierarchical dense recursive network for image super-resolution," *Pattern Recognit.*, vol. 107, Nov. 2020, Art. no. 107475.

[12] S. Chen, L. Zhang, and L. Zhang, "Cross-scope spatial–spectral information aggregation for hyperspectral image super-resolution," *IEEE Trans. Image Process.*, vol. 33, pp. 5878–5891, 2024.

[13] M. Protter, M. Elad, H. Takeda, and P. Milanfar, "Generalizing the nonlocal-means to super-resolution reconstruction," *IEEE Trans. Image Process.*, vol. 18, no. 1, pp. 36–51, Jan. 2009.

[14] S. D. Babacan, R. Molina, and A. K. Katsaggelos, "Variational Bayesian super resolution," *IEEE Trans. Image Process.*, vol. 20, no. 4, pp. 984–999, Apr. 2011.

[15] S. Chen, L. Zhang, and L. Zhang, "MSDformer: Multiscale deformable transformer for hyperspectral image super-resolution," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5525614.

[16] Z. Xiao, D. Kai, Y. Zhang, X. Sun, and Z. Xiong, "Asymmetric event-guided video super-resolution," in *Proc. ACM Int. Conf. Multimedia*, 2024, pp. 2409–2418.

[17] Y. Xu, L. Zhang, B. Du, and L. Zhang, "Hyperspectral anomaly detection based on machine learning: An overview," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 15, pp. 3351–3364, 2022.

[18] X. Wang, K. C. K. Chan, K. Yu, C. Dong, and C. C. Loy, "EDVR: Video restoration with enhanced deformable convolutional networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2019, p. 0.

[19] K. C. K. Chan, X. Wang, K. Yu, C. Dong, and C. C. Loy, "BasicVSR: The search for essential components in video super-resolution and beyond," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 4947–4956.

[20] T. Isobe, X. Jia, S. Gu, S. Li, S. Wang, and Q. Tian, "Video super-resolution with recurrent structure-detail network," in *Proc. Eur. Conf. Comput.*, Aug. 2020, pp. 645–660.

[21] Z. Xiao, D. Kai, Y. Zhang, Z.-J. Zha, X. Sun, and Z. Xiong, "Event-adapted video super-resolution," in *Proc. Eur. Conf. Comput. Vis.*, Oct. 2024, pp. 217–235.

[22] Y. Jo, S. W. Oh, J. Kang, and S. J. Kim, "Deep video super-resolution network using dynamic upsampling filters without explicit motion compensation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 3224–3232.

[23] J. Yu, J. Liu, L. Bo, and T. Mei, "Memory-augmented non-local attention for video super-resolution," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 17834–17843.

[24] Y. Xiao et al., "Local-global temporal difference learning for satellite video super-resolution," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 34, no. 4, pp. 2789–2802, Apr. 2024.

[25] M. S. M. Sajjadi, R. Vemulapalli, and M. Brown, "Frame-recurrent video super-resolution," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 6626–6634.

[26] T. Isobe, F. Zhu, X. Jia, and S. Wang, "Revisiting temporal modeling for video super-resolution," 2020, *arXiv:2008.05765*.

[27] K. C. K. Chan, S. Zhou, X. Xu, and C. C. Loy, "BasicVSR++: Improving video super-resolution with enhanced propagation and alignment," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 5972–5981.

[28] C. Liu, H. Yang, J. Fu, and X. Qian, "Learning trajectory-aware transformer for video super-resolution," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 5687–5696.

[29] A. Kappeler, S. Yoo, Q. Dai, and A. K. Katsaggelos, "Video super-resolution with convolutional neural networks," *IEEE Trans. Comput. Imag.*, vol. 2, no. 2, pp. 109–122, Jun. 2016.

[30] M. Haris, G. Shakhnarovich, and N. Ukita, "Recurrent back-projection network for video super-resolution," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 3897–3906.

[31] Y. Huang, W. Wang, and L. Wang, "Bidirectional recurrent convolutional networks for multi-frame super-resolution," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 28, Dec. 2015, pp. 235–243.

[32] E. Ilg, N. Mayer, T. Saikia, M. Keuper, A. Dosovitskiy, and T. Brox, "FlowNet 2.0: Evolution of optical flow estimation with deep networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2462–2470.

[33] A. Ranjan and M. J. Black, "Optical flow estimation using a spatial pyramid network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 4161–4170.

[34] J. Caballero et al., "Real-time video super-resolution with spatio-temporal networks and motion compensation," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, Jun. 2017, pp. 4778–4787.

[35] L. Wang, Y. Guo, L. Liu, Z. Lin, X. Deng, and W. An, "Deep video super-resolution using HR optical flow estimation," *IEEE Trans. Image Process.*, vol. 29, pp. 4323–4336, 2020.

[36] S. Shi, J. Gu, L. Xie, X. Wang, Y. Yang, and C. Dong, "Rethinking alignment in video super-resolution transformers," in *Proc. Adv. Neural Inf. Process. Syst.*, 2022, pp. 36081–36093.

[37] W. Wen, W. Ren, Y. Shi, Y. Nie, J. Zhang, and X. Cao, "Video super-resolution via a spatio-temporal alignment network," *IEEE Trans. Image Process.*, vol. 31, pp. 1761–1773, 2022.

[38] Y. Tian, Y. Zhang, Y. Fu, and C. Xu, "TDAN: Temporally-deformable alignment network for video super-resolution," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 3360–3369.

[39] P. Yi, Z. Wang, K. Jiang, J. Jiang, and J. Ma, "Progressive fusion video super-resolution network via exploiting non-local spatio-temporal correlations," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 3106–3115.

[40] Y. Luo, L. Zhou, S. Wang, and Z. Wang, "Video satellite imagery super resolution via convolutional neural networks," *IEEE Geosci. Remote Sens. Lett.*, vol. 14, no. 12, pp. 2398–2402, Dec. 2017.

[41] A. Xiao, Z. Wang, L. Wang, and Y. Ren, "Super-resolution for 'jilin-1' satellite video imagery via a convolutional network," *Sensors*, vol. 18, no. 4, p. 1194, 2018.

[42] K. Jiang, Z. Wang, P. Yi, and J. Jiang, "A progressively enhanced network for video satellite imagery superresolution," *IEEE Signal Process. Lett.*, vol. 25, no. 11, pp. 1630–1634, Nov. 2018.

[43] H. Liu, Y. Gu, T. Wang, and S. Li, "Satellite video super-resolution based on adaptively spatiotemporal neighbors and nonlocal similarity regularization," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 12, pp. 8372–8383, Dec. 2020.

[44] H. Liu and Y. Gu, "Deep joint estimation network for satellite video super-resolution with multiple degradations," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5621015.

[45] Z. He and D. He, "A unified network for arbitrary scale super-resolution of video satellite images," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 10, pp. 8812–8825, Oct. 2021.

[46] Y. Xiao, X. Su, Q. Yuan, D. Liu, H. Shen, and L. Zhang, "Satellite video super-resolution via multiscale deformable convolution alignment and temporal grouping projection," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5610819.

[47] N. Ni and L. Zhang, "Deformable convolution alignment and dynamic scale-aware network for continuous-scale satellite video super-resolution," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, 2024, Art. no. 5610017.

[48] S. Bako et al.., "Kernel-predicting convolutional networks for denoising Monte Carlo renderings," *ACM Trans. Graph.*, vol. 36, no. 4, pp. 1–14, Aug. 2017.

[49] B. Mildenhall, J. T. Barron, J. Chen, D. Sharlet, R. Ng, and R. Carroll, "Burst denoising with kernel prediction networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 2502–2510.

[50] Y. Jiang, B. Wronski, B. Mildenhall, J. T. Barron, Z. Wang, and T. Xue, "Fast and high quality image denoising via malleable convolution," in *Proc. Eur. Conf. Comput. Vis.*, 2022, pp. 429–446.

[51] Y. Chen, X. Dai, M. Liu, D. Chen, L. Yuan, and Z. Liu, "Dynamic convolution: Attention over convolution kernels," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 11030–11039.

[52] W. Li, X. Tao, T. Guo, L. Qi, J. Lu, and J. Jia, "MuCAN: Multi-correspondence aggregation network for video super-resolution," in *Proc. 16th Eur. Conf. Comput. Vis. (ECCV)*, Glasgow, U.K. Springer, Aug. 2020, pp. 335–351.

[53] T. Isobe et al.., "Video super-resolution with temporal group attention," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 8008–8017.

[54] J. Liang et al.., “Recurrent video restoration transformer with guided deformable attention,” in *Proc. Adv. Neural Inf. Process. Syst.*, 2022, pp. 378–393.

[55] J. Hu, L. Shen, and G. Sun, “Squeeze-and-excitation networks,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2018, pp. 7132–7141.

[56] X. Mao, Y. Liu, F. Liu, Q. Li, W. Shen, and Y. Wang, “Intriguing findings of frequency selection for image deblurring,” in *Proc. 37th AAAI Conf. Artif. Intell.*, 2023, vol. 37, no. 2, pp. 1905–1913.

[57] F. Li, L. Zhang, Z. Liu, J. Lei, and Z. Li, “Multi-frequency representation enhancement with privilege information for video super-resolution,” in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2023, pp. 12814–12825.

[58] K. Xu, Z. Yu, X. Wang, M. B. Mi, and A. Yao, “Enhancing video super-resolution via implicit resampling-based alignment,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2024, pp. 2546–2555.

[59] W. Lian and W. Lian, “Sliding window recurrent network for efficient video super-resolution,” in *Proc. Eur. Conf. Comput. Vis.*, Oct. 2022, pp. 591–601.

[60] B. Xia et al.., “Structured sparsity learning for efficient video super-resolution,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2023, pp. 22638–22647.

[61] W. Zhou, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, “Image quality assessment: From error visibility to structural similarity,” *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, Apr. 2004.

[62] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, “The unreasonable effectiveness of deep features as a perceptual metric,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 586–595.

[63] S. Nah et al.., “NTIRE 2019 challenge on video deblurring and super-resolution: Dataset and study,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2019, pp. 1–10.

[64] C. Liu and D. Sun, “On Bayesian adaptive video super resolution,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 2, pp. 346–360, Feb. 2014.

[65] J. Liang et al.., “VRT: A video restoration transformer,” *IEEE Trans. Image Process.*, vol. 33, pp. 2171–2182, 2024.

[66] Z. Qin, P. Zhang, F. Wu, and X. Li, “FcaNet: Frequency channel attention networks,” in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, Oct. 2021, pp. 783–792.

Yi Xiao (Graduate Student Member, IEEE) received the B.S. degree from the School of Mathematics and Physics, China University of Geosciences, Wuhan, China, in 2020. He is currently pursuing the Ph.D. degree with the School of Geodesy and Geomatics, Wuhan University, Wuhan.

His major research interests include remote sensing image/video processing, and computer vision. More details can be found at https://xy-boy.github.io/ .

Qiangqiang Yuan (Member, IEEE) received the B.S. degree in surveying and mapping engineering and the Ph.D. degree in photogrammetry and remote sensing from Wuhan University, Wuhan, China, in 2006 and 2012, respectively.

In 2012, he joined the School of Geodesy and Geomatics, Wuhan University, where he is currently a Professor. He has published more than 90 research articles, including more than 70 peer-reviewed articles in international journals, such as *Remote Sensing of Environment*, *ISPRS Journal of Photogrammetry and Remote Sensing*, IEEE Transactions on Image Processing, and IEEE Transactions on Geoscience and Remote Sensing. His research interests include image reconstruction, remote sensing image processing and application, and data fusion.

Dr. Yuan was a recipient of the Youth Talent Support Program of China in 2019, the Top-Ten Academic Star of Wuhan University in 2011, and the Recognition of Best Reviewers of IEEE Geoscience and Remote Sensing Letters in 2019. In 2014, he received Hong Kong Scholar Award from the Society of Hong Kong Scholars and China National Postdoctoral Council. He is an associate editor of five international journals and has frequently served as a referee for more than 40 international top journals, such as *Nature Climate Change* and *Nature Communications*.

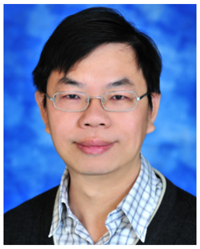
Kui Jiang (Member, IEEE) received the Ph.D. degree from the School of Computer Science, Wuhan University, Wuhan, China, in 2022.

In July 2023, he was a Chief Engineer with Huawei Technologies, Cloud BU, Hangzhou, China. He is currently an Associate Professor with the School of Computer Science and Technology, Harbin Institute of Technology, Harbin, China. His research interests include image/video processing and computer vision. He received the 2023 CSIG Excellent Doctoral Dissertation Award and the 2022 ACM Wuhan Doctoral Dissertation Award.

Yuzeng Chen received the B.S. degree in geographic information science from the Southwest University of Science and Technology, Mianyang, China, in 2020, and the M.S. degree in surveying and mapping engineering from Central South University, Changsha, China, in 2023. He is currently pursuing the Ph.D. degree with the School of Geodesy and Geomatics, Wuhan University, Wuhan.

His research interests include remote-sensing video processing and computer vision.

Shiqi Wang (Senior Member, IEEE) received the Ph.D. degree in computer application technology from Peking University in 2014. He is currently an Associate Professor with the Department of Computer Science, City University of Hong Kong. He has proposed more than 70 technical proposals to ISO/MPEG, ITU-T, and AVS standards. He has authored or co-authored more than 300 refereed journal articles/conference papers, including more than 100 IEEE Transactions. His research interests include video compression, image/video quality assessment, video coding for machine, and semantic communication. He received the Best Paper Award from IEEE VCIP 2019, ICME 2019, IEEE Multimedia 2018, and PCM 2017. His co-authored article received the Best Student Paper Award from IEEE ICIP 2018. He served or serves as an Associate Editor for IEEE Transactions on Image Processing, IEEE Transactions on Circuits and Systems for Video Technology, IEEE Transactions on Multimedia, IEEE Transactions on Cybernetics, IEEE Access, and *APSIPA Transactions on Signal and Information Processing*.



Chia-Wen Lin (Fellow, IEEE) received the Ph.D. degree in electrical engineering from National Tsing Hua University (NTHU), Hsinchu, Taiwan, in 2000.

He was with the Department of Computer Science and Information Engineering, National Chung Cheng University, Taiwan, from 2000 to 2007. Prior to joining academia, he was with the Information and Communications Research Laboratories, Industrial Technology Research Institute, Hsinchu, from 1992 to 2000. He is currently a Professor with the Department of Electrical Engineering and the Institute of Communications Engineering, NTHU. He is also the Deputy Director of the AI Research Center, NTHU. His research interests include image and video processing, computer vision, and video networking. From 2014 to 2015, he was a Steering Committee Member of IEEE Transactions on Multimedia. His articles received the Best Paper Award of IEEE VCIP 2015 and the Young Investigator Award of VCIP 2005. He received the Outstanding Electrical Professor Award presented by the Chinese Institute of Electrical Engineering in 2019 and the Young Investigator Award presented by the Ministry of Science and Technology, Taiwan, in 2006. He was the Chair of the Multimedia Systems and Applications Technical Committee of the IEEE Circuits and Systems Society from 2013 to 2015. He is the Chair of the Steering Committee of IEEE ICME. He has served as the Technical Program Co-Chair for IEEE ICME 2010, the General Co-Chair for IEEE VCIP 2018, and the Technical Program Co-Chair for IEEE ICIP 2019. He has served as an Associate Editor for IEEE Transactions on Image Processing, IEEE Transactions on Circuits and Systems for Video Technology, IEEE Transactions on Multimedia, IEEE Multimedia, and *Journal of Visual Communication and Image Representation*. He served as a Distinguished Lecturer for the IEEE Circuits and Systems Society from 2018 to 2019.