remote sensing

MDPI

Article

KAN-Enhanced Alignment and Fusion for Lightweight Satellite Video Super-Resolution

Junjie Xiong [1], Haopeng Zhang [1,2,3,*] and Zhiguo Jiang [2]

[1] School of Astronautics, Beihang University, Beijing 102206, China
[2] Tianmushan Laboratory, Beihang University, Hangzhou 311115, China
[3] Key Laboratory of Spacecraft Design Optimization and Dynamic Simulation Technology, Ministry of Education, Beijing 102206, China
* Correspondence: zhanghaopeng@buaa.edu.cn; Tel.: +86-10-6171-6978

Highlights

What are the main findings?

- A lightweight satellite video super-resolution method, termed KANSVSR, is proposed by incorporating Kolmogorov–Arnold Networks (KAN) into a conventional satellite video super-resolution (SVSR) framework, thereby improving cross-frame feature modeling.
- Two dedicated modules, namely the KAN Alignment Module (KAM) and the KAN Fusion Module (KFM), are designed to improve neighboring-frame representation and multi-frame feature aggregation, leading to enhanced reconstruction performance with a compact model.

What are the implications of the main findings?

- The results demonstrate that incorporating KAN-based learnable functions into video super-resolution can effectively enhance feature representation, offering a new modeling perspective for lightweight remote sensing video reconstruction.
- The proposed method achieves superior performance across multiple evaluation metrics while maintaining low model complexity, providing an efficient and effective solution for satellite video super-resolution.

Abstract

Satellite video super-resolution (SVSR) aims to reconstruct high-resolution video frames from low-resolution satellite observations, providing enhanced visual details for remote sensing applications. Despite recent progress, existing methods still suffer from limited alignment accuracy under complex motion and insufficient feature aggregation across frames, which restricts reconstruction quality. To address these issues, we propose a lightweight SVSR framework that incorporates Kolmogorov–Arnold Networks (KAN) into both the alignment and fusion processes. Specifically, a KAN-based spatial attention module is introduced to enhance the first-order and second-order neighboring frames, improving the accuracy of frame alignment. In addition, a KAN-based channel attention mechanism is adopted to facilitate more effective multi-frame feature aggregation. Benefiting from these designs, the proposed framework achieves strong reconstruction capability while maintaining a lightweight model structure. Extensive experiments demonstrate that the proposed method achieves superior performance in terms of PSNR, SSIM, LPIPS, and tOF compared with existing approaches, verifying the effectiveness of integrating KAN into SVSR.

Check for updates

Academic Editors: Xiaowen Liu and Yijun Chen

Received: 5 April 2026
Revised: 9 May 2026
Accepted: 14 May 2026
Published: 16 May 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

Keywords: satellite video super-resolution; Kolmogorov–Arnold networks; attention mechanism

1. Introduction

Satellite video has become an increasingly important data source in the field of remote sensing, providing continuous temporal observations for monitoring dynamic changes on the Earth's surface. Remote sensing data have been widely used in a variety of applications, including traffic monitoring [1,2], environmental surveillance [3,4], object detection [5,6], and target tracking [7,8]. Compared with traditional single-frame remote sensing imagery, satellite video enables more comprehensive analysis of object motion and scene evolution by capturing temporal dynamics. However, due to limitations in imaging sensors, acquisition conditions, and transmission bandwidth, the spatial resolution of satellite video is often significantly lower than that of static remote sensing images [9]. This limitation leads to blurred structural details and degraded visual quality, which can negatively affect the performance of downstream tasks [10]. Satellite video super-resolution (SVSR) has therefore emerged as an important technique to enhance the spatial resolution of satellite video sequences by reconstructing high-resolution frames from low-resolution observations [11]. Compared with single-image super-resolution methods [12,13], SVSR can further exploit temporal information across consecutive frames to improve reconstruction quality and robustness [14,15].

Recently, SVSR has attracted increasing attention. Many existing approaches are developed by adapting video super-resolution (VSR) techniques [16,17] originally designed for natural scenes to the remote sensing domain. These methods typically exploit temporal information across multiple frames and employ feature alignment and fusion strategies to enhance reconstruction quality [18]. Benefiting from the rapid development of deep learning, such approaches have achieved remarkable progress in video restoration tasks and provide an important foundation for SVSR research. However, due to the significant differences between satellite videos and natural videos in terms of imaging mechanisms, object scale distribution, and scene complexity, directly transferring natural-scene VSR architectures often fails to fully adapt to remote sensing scenarios, leaving several challenges to be addressed [19]. Despite the progress achieved by existing SVSR approaches, several limitations still remain. First, accurate frame alignment remains challenging in satellite video due to complex motion patterns and large spatial variations [20]. Most existing methods rely on optical-flow-based warping or deformable convolution to align neighboring frames. Although these techniques are effective in many scenarios, they mainly focus on geometric transformation and often lack sufficient capability to enhance informative features before alignment, which may limit the robustness of feature matching across frames. Second, effective multi-frame feature aggregation is still difficult [21,22]. Many SVSR frameworks employ relatively simple fusion strategies, such as convolution-based aggregation or basic attention mechanisms, which may not fully exploit the complementary information contained in neighboring frames [23]. These limitations motivate the exploration of more powerful feature modeling mechanisms to improve both frame alignment and multi-frame feature fusion in SVSR.

Recently, inspired by the Kolmogorov–Arnold representation theorem [24,25], Kolmogorov–Arnold Networks (KAN) [26] have attracted growing attention due to their strong nonlinear representation capability and flexible functional modeling. Unlike conventional neural networks that mainly rely on fixed activation functions, KAN models complex mappings through learnable univariate functions on network connections, providing

enhanced representation power and interpretability [27]. Compared with standard MLPs or convolutional transformations, KAN allows the nonlinear mapping function itself to be adaptively learned, rather than relying only on fixed activations, linear projections, or local convolutional operations. This property provides a flexible feature modeling mechanism for capturing complex spatial and channel relationships in visual restoration tasks. Although KAN has shown promising potential in several machine learning and vision tasks [28], its application to video restoration, particularly SVSR, remains largely unexplored. Motivated by its flexible feature modeling capability, we investigate the potential of KAN in SVSR and incorporate it into both frame alignment and feature aggregation processes to improve the overall reconstruction performance.

In this paper, we propose a lightweight SVSR framework by incorporating KAN into both frame alignment and feature aggregation. Specifically, a KAN spatial attention module is introduced in the alignment stage to enhance the first-order and second-order neighboring-frame features before alignment, enabling more reliable feature matching and improving alignment accuracy. Furthermore, a KAN channel attention mechanism is employed during the fusion process to model channel-wise dependencies among aggregated temporal features and facilitate more effective multi-frame information aggregation. In addition, the proposed framework adopts a compact propagation structure, allowing the model to achieve a strong reconstruction performance while maintaining a lightweight design. Through these designs, our method effectively improves both alignment robustness and feature fusion capability for SVSR. The source code will be available at: https://github.com/DylanFantast/KANSVSR (accessed on 1 April 2026).

The main contributions of this paper can be summarized as follows:

(1) We propose a KAN alignment module (KAM) that introduces KAN spatial attention to enhance neighboring-frame feature representations, improving alignment accuracy under complex motion in satellite videos.

(2) We design a KAN fusion module (KFM) that introduces KAN channel attention to model channel-wise dependencies among multi-frame features, enhancing the robustness of spatiotemporal feature aggregation.

(3) We develop a lightweight SVSR framework that achieves strong reconstruction performance while maintaining a compact model structure and high computational efficiency.

The remainder of this paper is organized as follows. Section 2 reviews related work on SVSR and KAN. Section 3 presents the overall architecture of the proposed KANSVSR framework and describes the key modules in detail. Section 4 introduces the experimental settings. Section 5 presents the results and discussion, including both quantitative and qualitative analyses. Finally, Section 6 concludes the paper.

2. Related Works

2.1. Super-Resolution for Satellite Videos

VSR aims to reconstruct high-resolution frames from low-resolution video sequences by exploiting both spatial information within individual frames and temporal correlations across multiple frames [29]. Early VSR methods were mainly based on traditional optimization frameworks. For example, Schultz and Stevenson [30] proposed a motion-compensated subsampling model with affine motion estimation, while Liu and Sun [31] introduced a Bayesian formulation to jointly estimate motion, blur, and noise. Ma et al. [32] further developed an expectation–maximization (EM) framework for residual blur estimation and high-resolution reconstruction. Although these methods provided important insights, their performance was limited by inaccurate motion estimation and high computational cost.

With the development of deep learning, convolutional neural networks (CNNs) have become the dominant paradigm for VSR. Early works such as SRCNN [33] extended single-

image super-resolution to video scenarios, followed by a series of CNN-based methods. Existing deep learning-based VSR approaches can be broadly categorized into two groups: methods without explicit alignment and methods with alignment.

Methods without alignment avoid explicit motion compensation by directly modeling temporal relationships. For instance, DUF [34] dynamically generates upsampling filters conditioned on input frames, while RBPN [35] employs a recurrent encoder–decoder architecture to integrate spatial and temporal information across frames. These approaches simplify the pipeline but may struggle to handle complex motion accurately.

In contrast, alignment-based methods explicitly model motion between frames and generally achieve better performance. TDAN [36] introduces deformable convolution to implicitly align neighboring frames with the reference frame. EDVR [37] further improves alignment accuracy by adopting pyramid, cascading, and deformable convolutions. BasicVSR [38] and IconVSR [38] propose a unified framework with bidirectional propagation to efficiently exploit temporal information, while BasicVSR++ [39] enhances both propagation and alignment for improved reconstruction quality. More recently, RVRT [40] incorporates Transformer-based architectures to capture long-range temporal dependencies, providing stronger modeling capability for complex motion scenarios. These methods establish a solid foundation for subsequent research in SVSR.

In the remote sensing domain, researchers have gradually extended VSR techniques to SVSR. Zhang et al. [41] first adapted deep VSR models to satellite imagery, paving the way for subsequent research in this area. To address the unique characteristics of remote sensing data, such as large-scale scenes, small objects, and complex motion patterns, various SVSR methods have been proposed. Many works build upon alignment-based frameworks. For example, Xiao et al. [42] proposed a deformable-convolution-based alignment strategy to better handle sparse motion and multi-scale objects. Wei et al. [43] employed a unidirectional recurrent network for VSR and introduced an auxiliary network independent of the super-resolution framework to simulate realistic degradations. Wang et al. [44] designed a lightweight temporal fusion mechanism to improve efficiency, while Xiao et al. [45] introduced a multi-axis feature diversity enhancement module to strengthen feature representation in both spatial and channel dimensions.

In addition to alignment-based methods, several studies explore alternative architectures. Liu and Gu [46] proposed a dual-stream network, where one branch predicts the high-resolution image and the other estimates the blur kernel, with cross-task feature fusion to enhance reconstruction. Shen et al. [47] incorporated an edge prediction branch into the EDVR framework, enabling joint optimization of structural information and image reconstruction.

Moreover, some approaches avoid explicit alignment by directly modeling spatio-temporal features. For instance, 3D convolution-based methods [48,49] perform feature extraction and fusion in a unified manner without motion estimation. These methods simplify the pipeline but may be less effective in handling complex motion.

Beyond supervised learning, unsupervised SVSR approaches have also been explored. For example, methods based on cycle consistency or degradation modeling [50] employ paired downsampling and upsampling networks, eliminating the need for LR-HR training pairs. In addition, GAN-based methods [51] introduce adversarial learning and attention mechanisms to enhance perceptual quality and texture realism.

Despite these advances, several challenges remain in SVSR. First, most existing methods primarily focus on geometric alignment while paying limited attention to enhancing the feature representation of neighboring frames, which may degrade alignment accuracy under complex motion or large-scale variations. Second, multi-frame feature aggregation is often implemented using relatively simple strategies, limiting the effective utilization of

complementary temporal information. Third, some approaches rely on complex network architectures, resulting in high computational cost and reduced practicality in real-world remote sensing applications. Therefore, it is still necessary to develop SVSR methods that enhance feature representation, improve temporal fusion, and maintain a lightweight design.

2.2. Kolmogorov–Arnold Networks

Kolmogorov–Arnold Networks (KAN) [26] are a recently proposed neural network architecture inspired by the Kolmogorov–Arnold representation theorem [24,25]. By introducing learnable univariate functions on network connections, KAN enhances the nonlinear function approximation capability of neural networks while improving interpretability and parameter efficiency [52]. Due to its strength in modeling continuous functions, KAN was initially applied to time-series prediction and signal modeling tasks.

For example, Vaca-Rubio et al. [53] applied KAN to satellite traffic time-series forecasting, demonstrating its effectiveness in capturing temporal dependencies. In addition, KAN has been used in signal super-resolution tasks such as inertial signal enhancement, where spline-based functions enable more accurate modeling of continuous dynamics and frequency components compared with conventional discrete networks [54].

Subsequently, KAN has been introduced into computer vision and has shown promising results in tasks such as image segmentation, image reconstruction, and super-resolution. For instance, U-KAN [55] integrates KAN modules into the U-Net architecture for efficient medical image segmentation. In face super-resolution, KAN-based hybrid attention mechanisms have been employed to improve nonlinear feature modeling and enhance fine-detail reconstruction [56]. Furthermore, KAN has also been explored in high-resolution imaging tasks such as ultrasound localization microscopy, where it achieves improved reconstruction accuracy with a compact model structure [57]. These studies demonstrate the effectiveness of KAN in capturing complex spatial patterns and enhancing feature representation.

In the remote sensing domain, KAN has been further applied to image super-resolution and spectral reconstruction tasks. For example, HyperKAN [58] models continuous spectral variations using spline-based functions, leading to improved spectral reconstruction in hyperspectral image super-resolution. KMSR [59] combines CNN and KAN to enhance nonlinear spectral–spatial representation, achieving better spectral consistency and spatial detail recovery. In addition, KAN has been integrated with adversarial learning for hyperspectral image super-resolution, further improving both spatial fidelity and spectral accuracy [60]. These works highlight the strong potential of KAN in modeling complex relationships in high-dimensional remote sensing data.

Despite these encouraging results, the application of KAN to SVSR remains largely unexplored. Existing KAN-based vision and remote sensing studies mainly focus on single-image analysis, segmentation, or spectral–spatial reconstruction, while the potential of KAN for temporal feature interaction in video restoration has not been sufficiently investigated. In particular, how to leverage the nonlinear modeling capability of KAN to enhance neighboring-frame feature representation and improve cross-frame information modeling for video alignment and multi-frame fusion is still an open problem.

3. Methods

3.1. Overall Architecture

Figure 1 illustrates the overall architecture of the proposed KANSVSR network. The model follows a bidirectional recurrent framework for video super-resolution, where temporal information is propagated along the sequence to exploit inter-frame correlations. Given a low-resolution video sequence $x_1, x_2, \ldots, x_T$, where $x_t \in \mathbb{R}^{3\times H\times W}$ denotes the

t-th input frame, the goal is to reconstruct the corresponding high-resolution sequence $y_1, y_2, \ldots, y_T$, where $y_t \in \mathbb{R}^{3\times sH\times sW}$ and s represents the upscaling factor.

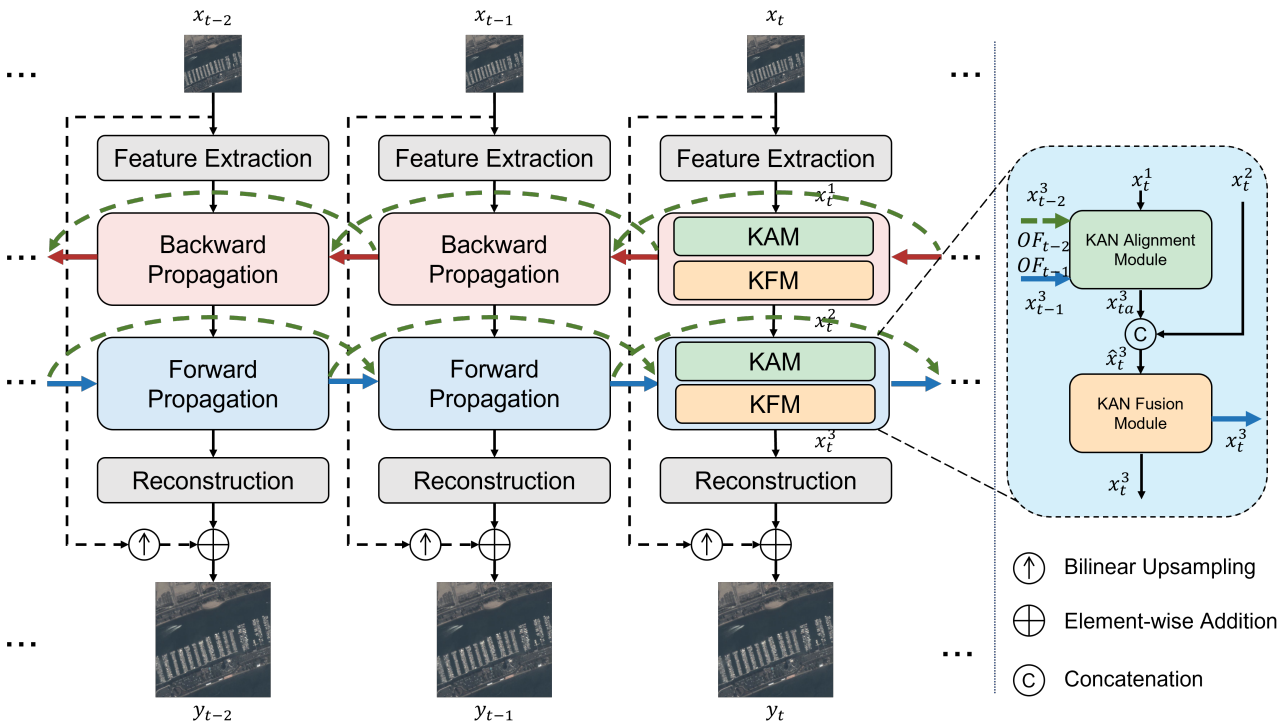

Figure 1. Architecture of the proposed KANSVSR network. Given a low-resolution input sequence $x_1, \ldots, x_T$, the network propagates temporal information through backward and forward branches to reconstruct the high-resolution output sequence $y_1, \ldots, y_T$. At each time step, the KAN Alignment Module (KAM) performs feature alignment guided by optical flow (OF), while the KAN Fusion Module (KFM) aggregates the aligned feature and the current propagated feature. The right part illustrates the detailed input–output relationship of KAM and KFM at time step t. The red and blue arrows indicate first-order backward and forward propagation, respectively, while the green dashed arrows indicate second-order propagation.

The overall framework mainly consists of three components: a feature extraction module, bidirectional propagation modules, and a reconstruction module. The feature extraction module first maps each input frame x_t into a shallow feature representation $x_t^1 \in \mathbb{R}^{C\times H\times W}$ using several convolutional and residual layers. These shallow features provide spatial representations of the current frame and serve as reference features for subsequent alignment and reconstruction.

Following recurrent video super-resolution frameworks such as BasicVSR++, the bidirectional propagation stage contains a backward branch and a forward branch that transmit information along the temporal dimension. For each time step t, the proposed KAM is used to enhance and align the neighboring-frame features before feature fusion. As illustrated in the right part of Figure 1, KAM takes the current shallow feature x_t^1, the neighboring propagated features x_{t-1}^3 and x_{t-2}^3, and the corresponding optical flows OF_{t-1} and OF_{t-2} as inputs, and produces the aligned feature x_{ta}^3:

$$x_{ta}^3 = \mathrm{KAM}(x_t^1, x_{t-1}^3, x_{t-2}^3, OF_{t-1}, OF_{t-2}). \tag{1}$$

Here, OF_{t-1} and OF_{t-2} denote the optical flows that guide the transformation of the neighboring features from time steps $t-1$ and $t-2$ to the current time step t, respectively. By incorporating the current-frame feature and optical-flow-guided neighboring features,

KAM enhances the representation of adjacent frames and generates a more reliable aligned feature for subsequent fusion.

The aligned feature x_{ta}^3 is then concatenated with the current propagated feature x_t^2 to form an intermediate feature $\hat{x}_t^3$, which is further fed into KFM for multi-frame feature aggregation:

$$\hat{x}_t^3 = \text{concat}(x_{ta}^3, x_t^2), \tag{2}$$

$$x_t^3 = \text{KFM}(\hat{x}_t^3). \tag{3}$$

Through the KAN-based channel attention mechanism, KFM adaptively recalibrates the importance of different feature channels and generates the updated propagated feature x_t^3. Finally, in the reconstruction stage, the fused features are passed through convolutional layers and upsampling operators to generate the high-resolution output frame y_t.

By introducing KAN-based attention mechanisms into both the alignment and fusion stages, the proposed KANSVSR enhances cross-frame feature modeling while maintaining a lightweight architecture, leading to improved reconstruction performance for satellite video super-resolution.

3.2. KAN Layer

KAN is a recently proposed neural architecture inspired by the Kolmogorov–Arnold representation theorem, which states that multivariate functions can be represented as compositions of univariate functions. Unlike conventional neural networks that apply fixed activation functions at nodes, KAN introduces learnable nonlinear functions on network edges, enabling more flexible function approximation.

In a KAN layer, each node is fully connected to all nodes in the subsequent layer, and each connection is associated with a learnable activation function. The output of a node is obtained by summing the responses from all incoming edges. Specifically, the learnable function on each edge is modeled as a combination of a residual function and a weighted sum of B-spline basis functions:

$$f_{\text{KAN}}(x) = w_1 \cdot \text{SiLU}(x) + w_2 \cdot \sum_{i=0}^{G+k-1} c_i B_i(x), \tag{4}$$

where w_1 and w_2 are learnable weights, c_i are spline coefficients, and $B_i(x)$ denotes the B-spline basis functions.

The B-spline basis functions are defined over a set of knot points and constructed recursively. Given a spline order k and a grid size G, the knot vector is defined as

$$\mathcal{T} = (t_{-k}, \ldots, t_0, \ldots, t_G, \ldots, t_{G+k}), \tag{5}$$

where the knot points are uniformly spaced over the input domain. The zeroth-order basis functions are defined as

$$B_i^0(x) = \begin{cases} 1, & t_i \le x < t_{i+1}, \\ 0, & \text{otherwise}, \end{cases} \tag{6}$$

and for $k > 0$:

$$B_i^k(x) = \frac{x - t_i}{t_{i+k} - t_i} B_i^{k-1}(x) + \frac{t_{i+k+1} - x}{t_{i+k+1} - t_{i+1}} B_{i+1}^{k-1}(x). \tag{7}$$

Here, the spline order k controls the degree of the B-spline, while the grid size G determines the number of intervals used to partition the input domain, jointly defining the resolution and flexibility of the spline representation. Since the spline coefficients can adapt locally across different regions, KAN can learn input-dependent nonlinear transformations

on network edges. This differs from standard MLPs or convolutional transformations, which mainly rely on fixed activation functions, linear projections, or local convolutional operations. Therefore, KAN provides a more flexible feature modeling mechanism for capturing spatial and channel relationships in visual restoration tasks, motivating us to incorporate it into the alignment and fusion modules of the proposed SVSR network.

3.3. KAN Alignment Module

To improve the stability and accuracy of cross-frame feature alignment, we design the KAN alignment module (KAM). The module introduces KAN spatial attention (KSA) before the alignment process to enhance neighboring-frame features and strengthen the representation of important spatial regions. In this way, the network can adaptively emphasize informative spatial structures in adjacent frames before performing feature alignment, which improves the reliability of the subsequent alignment process.

As shown in Figure 2, at time step t, for the forward propagation branch, we first enhance the neighboring features x^3_{t-1} and x^3_{t-2} using KSA:

$$\tilde{x}^3_{t-i} = \mathrm{KSA}(x^3_{t-i}), \quad i \in \{1, 2\}. \tag{8}$$

Here, $\mathrm{KSA}(\cdot)$ denotes the KAN spatial attention module, which learns spatial weight maps to reweight the input features, thereby highlighting spatial regions that are more important for reconstructing the current frame. Specifically, KSA first extracts spatial contextual information through convolution operations with different receptive fields and generates spatial response maps via depthwise convolution. A KAN layer is then introduced to model the nonlinear mapping from spatial responses to attention weights, allowing the network to exploit the flexible representation capability of spline-based functions to refine the spatial attention. The resulting spatial weight map is applied to the input features to obtain enhanced neighboring features.

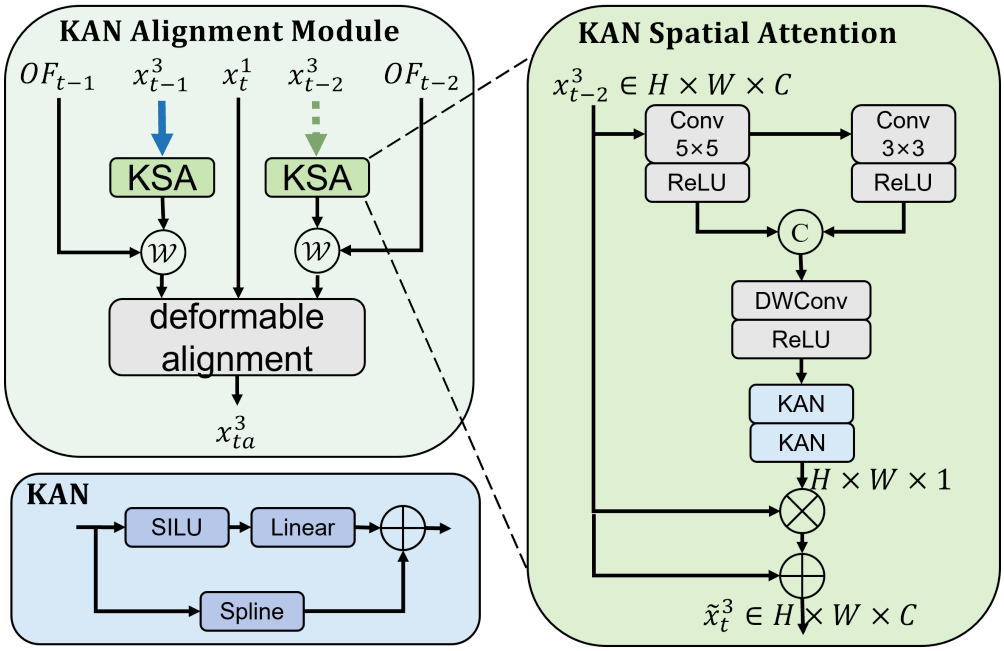

Figure 2. Detailed structure of the KAN Alignment Module (KAM). Neighboring-frame features are enhanced by KAN Spatial Attention (KSA) and then aligned through optical-flow-guided warping $\mathcal{W}$ and deformable alignment to obtain x^3_{ta}. In KSA, Conv, DWConv, and ReLU refer to convolution, depthwise convolution, and rectified linear unit, respectively. The blue arrow and green dashed arrow denote first- and second-order propagation. The circles labeled w and C denote warping and concatenation, while $\otimes$ and $\oplus$ denote element-wise multiplication and residual addition.

After obtaining the enhanced neighboring features, KAM first transforms them into the coordinate space of the current frame using optical-flow-guided warping, and then adopts flow-guided deformable alignment to further refine the alignment results:

$$x_{ta}^{3} = \mathrm{Align}\big(x_{t}^{1}, \mathcal{W}(\tilde{x}_{t-1}^{3}, OF_{t-1}), \mathcal{W}(\tilde{x}_{t-2}^{3}, OF_{t-2})\big). \tag{9}$$

Here, x_t^1 serves as the current-frame reference feature for deformable alignment. $\mathcal{W}(\cdot)$ denotes the optical-flow-guided warping operation, which initially transforms neighboring features into the current-frame coordinate space according to the optical flow. OF_{t-i} denotes the optical flow that guides the transformation of the neighboring feature from time step $t-i$ to the current time step t. $\mathrm{Align}(\cdot)$ denotes the flow-guided deformable alignment operation, which further models local residual offsets after the initial warping and outputs the final aligned feature x_{ta}^3.

By introducing KAN spatial attention before the alignment stage, KAM enhances the representation of important spatial information in neighboring frames, making the subsequent flow-guided deformable alignment more stable and reliable, thereby improving the quality of cross-frame feature alignment.

3.4. KAN Fusion Module

To further enhance temporal feature integration, we design the KAN fusion module (KFM) to perform channel-wise feature fusion. As illustrated in Figure 3, the aligned feature x_{ta}^3 produced by KAM is first concatenated with the current propagated feature x_t^2 to form the intermediate feature $\hat{x}_t^3 \in \mathbb{R}^{2C\times H\times W}$. Then, $\hat{x}_t^3$ is passed through residual blocks, where it is compressed from $2C$ to C channels and further refined to obtain $x_{tr}^3 \in \mathbb{R}^{C\times H\times W}$ for subsequent channel attention.

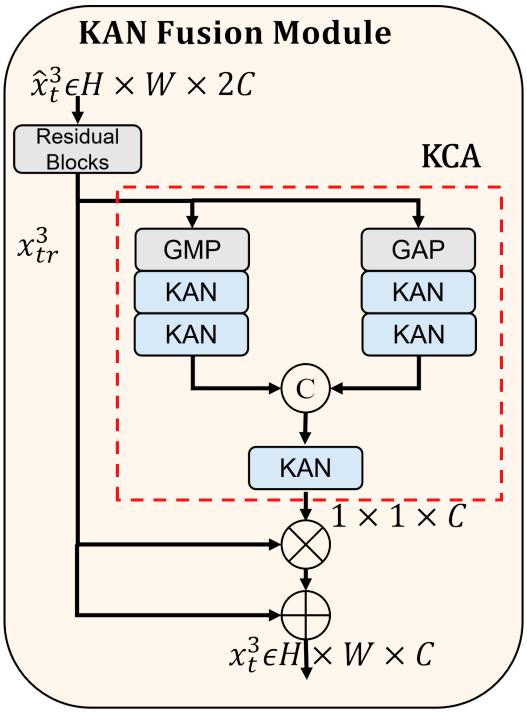

Figure 3. Detailed structure of the KAN Fusion Module (KFM). The intermediate feature $\hat{x}_t^3$ is compressed and refined by residual blocks, and then processed by KAN Channel Attention (KCA) to generate the channel attention vector. In KCA, GMP and GAP refer to global max pooling and global average pooling, respectively. The red dashed box highlights the KCA branch, where C denotes concatenation, and $\otimes$ and $\oplus$ denote element-wise multiplication and residual addition.

After feature refinement, a KAN channel attention (KCA) mechanism is employed to adaptively adjust the importance of different feature channels. Specifically, the refined feature is first aggregated along the spatial dimension using global max pooling (GMP) and global average pooling (GAP), generating two channel descriptors. These descriptors are then passed through two KAN layers to model nonlinear channel-wise dependencies. The resulting features are concatenated and further fused through an additional KAN layer to produce the final channel attention vector.

The channel attention vector is used to reweight the input feature along the channel dimension, followed by a residual connection to produce the fused feature:

$$x_t^3 = x_{tr}^3 + x_{tr}^3 \odot \mathrm{KCA}(x_{tr}^3), \tag{10}$$

where $\mathbf{x}_{tr}^3$ denotes the refined feature after residual blocks, $\mathrm{KCA}(\cdot)$ represents the KAN channel attention function, and $\odot$ denotes element-wise multiplication. By leveraging KAN to model channel relationships, the proposed KFM enables more flexible channel interaction and improves multi-frame feature fusion.

4. Experiments

4.1. Dataset

To evaluate the proposed method, we construct a satellite video super-resolution dataset by combining two widely used public datasets, namely SATSOT [61] and SAT-MTB [62]. These datasets are collected from multiple commercial satellite platforms, including Jilin-1, Skybox, and Carbonite-2, with a spatial resolution of approximately 1 m.

The constructed dataset covers a variety of ground objects commonly observed in remote sensing applications, such as vehicles, trains, ships, and aircraft. In addition, it includes diverse motion patterns, including high-speed translation, abrupt direction changes, occlusion, scale variation, and targets entering or leaving the field of view, providing challenging scenarios for SVSR evaluation.

To ensure consistency across samples, all video frames are uniformly cropped and resized to a spatial resolution of 640×640 pixels. In total, the dataset contains 172 video clips, each consisting of 100 consecutive frames. Among them, 150 clips are used for training and the remaining 22 clips are used for validation. The dataset is split at the video-clip level, and no frames from the same clip appear in both the training and validation sets, ensuring independence between the two sets.

The original videos are treated as high-resolution (HR) ground truth, while the corresponding low-resolution (LR) inputs are generated using a standard $4\times$ bicubic downsampling operation, resulting in frames of size 160×160. Some representative examples of the dataset are shown in Figure 4, illustrating the diversity of object categories and motion conditions.

4.2. Implementation Details

In the proposed KANSVSR framework, the number of feature channels in intermediate layers is set to 64, which provides a good balance between representation capacity and computational efficiency. For the embedded KAN modules, the spline order is set to 3 and the grid size is set to 5, enabling effective nonlinear modeling while maintaining a lightweight structure.

During training, sequences of 30 consecutive low-resolution frames are used as input, and supervision is applied to all corresponding high-resolution frames. To improve training efficiency, HR frames are randomly cropped into patches of size 256×256, and the corresponding LR patches are 64×64. Data augmentation techniques, including random

rotation as well as horizontal and vertical flipping, are also applied. During evaluation, all compared methods are applied to the complete validation video clips to reconstruct high-resolution frames for all time steps.

We adopt the Adam optimizer [63] with an initial learning rate of 1×10^{-4}, where $\beta_1 = 0.9$ and $\beta_2 = 0.99$. The learning rate is scheduled using the Cosine Annealing Warm Restart strategy [64], with a minimum learning rate of 1×10^{-7}. The total number of training iterations is set to 200,000. As the training objective, we employ the Charbonnier loss [65], which is a differentiable variant of the ℓ_1 loss and is more robust to outliers while better preserving image details. All experiments are implemented using the PyTorch 1.13.1 (CUDA 11.6) framework and conducted on two NVIDIA RTX 4090 GPUs.

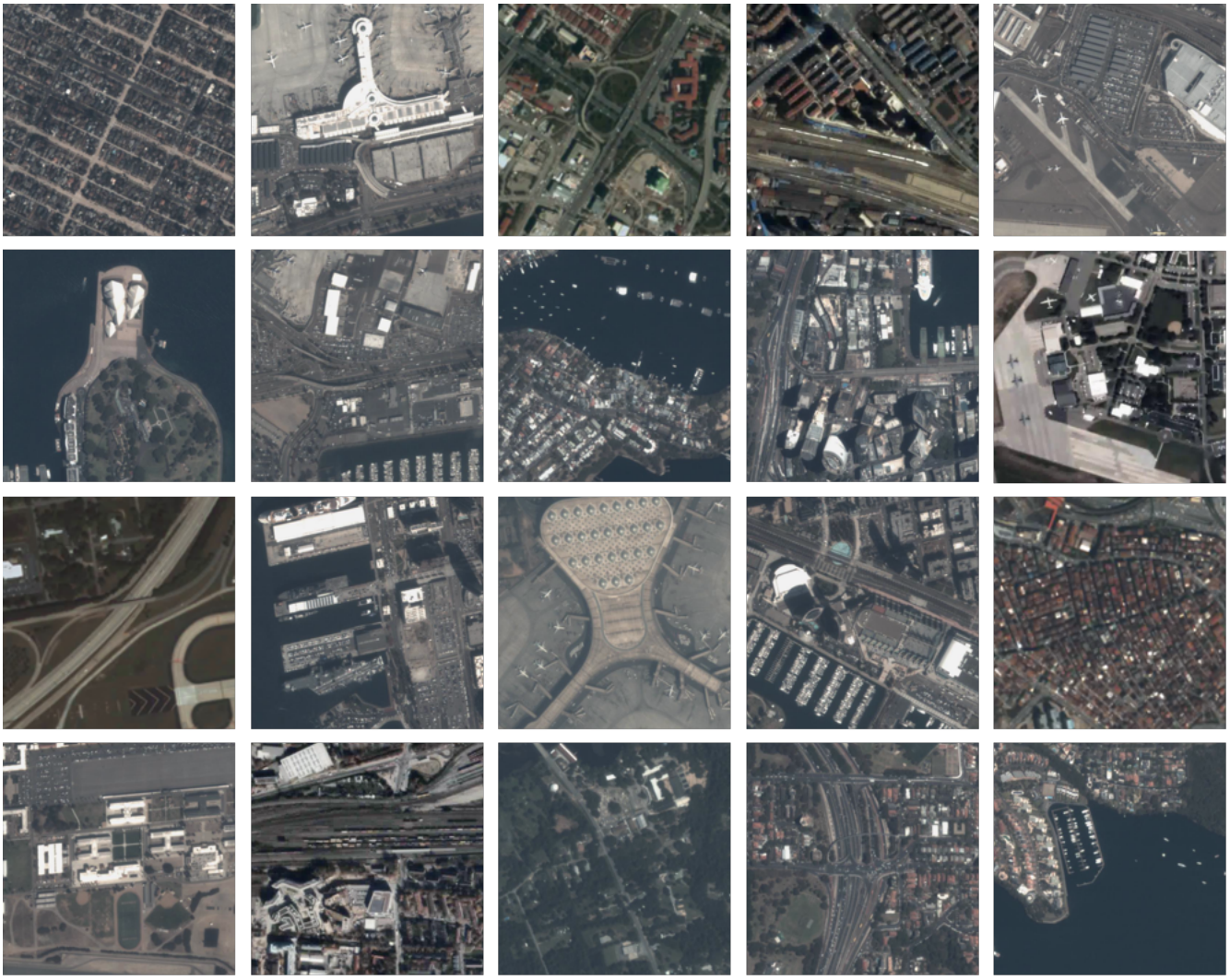

Figure 4. Examples from the constructed satellite video super-resolution dataset. The dataset includes various object types such as trains, cars, airplanes, and ships under diverse motion conditions including high-speed motion, turning, occlusion, and entering or exiting the field of view.

4.3. Evaluation Metrics

To quantitatively evaluate the reconstruction performance, we adopt four metrics, including Peak Signal-to-Noise Ratio (PSNR), Structural Similarity Index Measure (SSIM), Learned Perceptual Image Patch Similarity (LPIPS) [66], and temporal optical flow consistency (tOF) [67]. PSNR, SSIM, and LPIPS are used to measure pixel-level fidelity, structural similarity, and perceptual similarity in the feature space, respectively, while tOF is adopted to evaluate temporal consistency across reconstructed frames. A detailed description of

these metrics is provided in Table 1. PSNR, SSIM, and LPIPS are calculated on RGB images without boundary cropping, and the same evaluation setting is applied to all compared methods. For tOF, the same optical flow estimator is used for all methods to ensure a fair temporal consistency comparison. In addition to reconstruction quality and temporal consistency, we also report model parameters and FLOPs to assess computational complexity and efficiency. Together, these metrics provide a comprehensive evaluation of reconstruction performance, temporal stability, and resource consumption.

4.4. Compared Methods

To comprehensively evaluate the performance of the proposed KANSVSR model, we compare it with several representative video super-resolution methods, including both traditional interpolation-based approaches and state-of-the-art deep learning-based models.

First, we include the widely used bicubic interpolation as a baseline method, which serves as a non-learning reference for evaluating the effectiveness of learning-based approaches.

For deep learning-based methods, we consider several representative models originally developed for natural video super-resolution. EDVR [37] is a pioneering framework that introduces deformable convolution for feature alignment and has been widely adopted as a strong baseline. BasicVSR [38] and IconVSR [38] leverage bidirectional propagation and information reuse mechanisms to enhance temporal modeling. BasicVSR++ [39] further improves performance through enhanced propagation and alignment strategies, representing a strong baseline in CNN-based video super-resolution.

In addition, we include RVRT [40], a Transformer-based method that incorporates recurrent mechanisms to capture long-range temporal dependencies, providing powerful modeling capacity for complex motion scenarios.

To better reflect the performance of methods in remote sensing applications, we also compare recent approaches specifically designed for satellite video super-resolution. RASVSR [44] is a lightweight model that focuses on efficient alignment and reconstruction for satellite video, while MADNet [45] introduces multi-attention mechanisms to improve feature aggregation across frames.

For fair comparison, all learning-based baseline methods are retrained on the same training set using their official implementations. During evaluation, all methods are tested on the same validation clips with the same LR inputs and HR references. These compared methods cover a wide range of design paradigms, including traditional interpolation, CNN-based models, Transformer-based architectures, and satellite-specific approaches. Having such a diverse set of baselines ensures a fair and comprehensive evaluation of the proposed method under different modeling strategies and application scenarios.

Table 1. Table of detailed evaluation metrics.

Metric	Equation	Parameters	Description
Peak Signal-to-Noise Ratio (PSNR)	$MSE = \frac{1}{N}\sum_{i=1}^{N}\|X(i) - X_{SR}(i)\|_2^2$ $PSNR = 10\log_{10}\left(\frac{L^2}{MSE}\right)$	X is the target image; X_{SR} is the reconstructed image; L is the maximum pixel value of the image; N is the total number of pixels.	Measures pixel-level reconstruction fidelity between the reconstructed image and the reference image. A higher PSNR value indicates that the reconstructed result is closer to the ground truth.
Structural Similarity Index Measure (SSIM)	$SSIM(X, X_{SR}) = \frac{(2\mu_X\mu_{X_{SR}} + C_1)(2\sigma_{XX_{SR}} + C_2)}{(\mu_X^2 + \mu_{X_{SR}}^2 + C_1)(\sigma_X^2 + \sigma_{X_{SR}}^2 + C_2)}$	X is the target image; X_{SR} is the reconstructed image; μ_X and $\mu_{X_{SR}}$ are the mean values of the two images; σ_X and $\sigma_{X_{SR}}$ are the standard deviations; $\sigma_{XX_{SR}}$ denotes the covariance; C_1 and C_2 are constants for numerical stability.	Evaluates the structural similarity between the reconstructed image and the reference image. A higher SSIM value indicates better preservation of structural and textural information.
Learned Perceptual Image Patch Similarity (LPIPS)	$LPIPS(X, X_{SR}) = \sum_l w_l\|\phi_l(X) - \phi_l(X_{SR})\|_2^2$	$\phi_l(\cdot)$ denotes the deep feature extracted from the l-th layer; w_l is the corresponding layer weight; X is the target image; X_{SR} is the reconstructed image.	Measures perceptual similarity between the reconstructed image and the reference image in the deep feature space. A lower LPIPS value indicates better perceptual quality and closer visual similarity to the ground truth.
Temporal Optical Flow Consistency (tOF)	$\text{tOF} = \frac{1}{T-1}\frac{1}{N}\cdot \sum_{t=2}^{T}\left\|\mathcal{F}(X_{SR}^{t-1}, X_{SR}^{t}) - \mathcal{F}(X^{t-1}, X^{t})\right\|_1$	$\mathcal{F}(\cdot)$ denotes the optical flow estimator; X^t is the target image at time step t; X_{SR}^t is the reconstructed image at time step t; T is the number of frames; N is the number of evaluated pixels.	Measures the temporal motion consistency between the reconstructed video and the reference video. A lower tOF value indicates better temporal consistency.

5. Results and Discussion

5.1. Comparison with State-of-the-Art Methods

5.1.1. Frame Reconstruction Quality

Table 2 reports the frame reconstruction performance of different methods in terms of PSNR, SSIM, and LPIPS, as well as the model parameters. As can be observed, the proposed KANSVSR achieves the best overall reconstruction performance. Specifically, it obtains the highest PSNR (39.5676 dB) and SSIM (0.9623), while also achieving the lowest LPIPS score, indicating superior reconstruction fidelity, structural consistency, and perceptual quality.

Table 2. Comparison with Other Methods.

Method	Bicubic	BasicVSR	IconVSR	BasicVSR++	RVRT	RASVSR	MADNet	KANSVSR
PSNR (dB) ↑	35.9690	38.8943	38.7345	38.8152	38.8078	39.0090	**39.1264**	**39.5676**
SSIM ↑	0.9277	0.9558	0.9569	0.9580	0.9576	0.9589	**0.9607**	**0.9623**
LPIPS ↓	0.1868	0.0967	0.0863	0.0807	0.0812	0.0789	**0.0782**	**0.0777**
Params (M) ↓	/	6.3	8.7	7.3	6.2	**5.4**	6.7	**5.6**
Publications	/	CVPR'21	CVPR'21	CVPR'22	NIPS'22	JSTARS'24	TIP'25	Ours

Note: Metric arrows indicate the preferred direction: ↑ means higher is better, while ↓ means lower is better. The best and second-best results are highlighted in bold red and bold blue, respectively.

From the perspective of pixel-level reconstruction accuracy, the improvement in PSNR demonstrates that KANSVSR is able to more precisely recover fine-grained details from low-resolution inputs. This is particularly important for satellite imagery, where many targets, such as vehicles and ships, occupy only a limited number of pixels. Improved pixel fidelity therefore contributes to more accurate reconstruction of small-scale objects.

In terms of structural preservation, the higher SSIM achieved by KANSVSR indicates that the proposed method is more effective at maintaining the geometric structures and spatial relationships within the scene. This is critical for remote sensing applications, where the integrity of structures such as roads, buildings, and transportation networks directly affects downstream tasks, including object detection and scene understanding.

Regarding perceptual quality, the lower LPIPS score suggests that KANSVSR generates more visually realistic and sharper reconstructions. This improvement can be attributed to the enhanced feature representation enabled by the proposed KAN-based modules, which allow the network to better capture complex textures and spatial variations.

In addition to reconstruction quality, KANSVSR maintains a lightweight model with only 5.6 M parameters. Compared with existing methods, it achieves superior performance without introducing excessive model complexity, demonstrating a favorable balance between accuracy and model size.

The above improvements can be mainly attributed to the proposed KAN-based alignment and fusion mechanisms. The KAM enhances the representation of neighboring-frame features and improves alignment accuracy, while the KFM facilitates more effective cross-frame feature aggregation. These two components jointly enable better exploitation of temporal information, leading to improved reconstruction performance.

Overall, the quantitative results demonstrate that the proposed KANSVSR not only outperforms existing methods in terms of reconstruction quality but also maintains a compact model structure, making it suitable for practical satellite video super-resolution applications.

5.1.2. Temporal Consistency Analysis

To further evaluate the temporal coherence of reconstructed satellite videos, we compare different methods using tOF. Unlike frame-level metrics such as PSNR, SSIM, and

LPIPS, tOF measures whether the inter-frame motion of the reconstructed sequence is consistent with that of the HR reference sequence. A lower tOF value indicates better temporal consistency.

As shown in Table 3, KANSVSR achieves the lowest tOF among the compared methods, indicating that the reconstructed sequence better preserves temporal motion consistency with the HR reference. Compared with Bicubic interpolation, learning-based methods significantly reduce tOF values, showing the importance of temporal modeling in SVSR. Among the learning-based methods, BasicVSR++, RASVSR, and MADNet obtain relatively close tOF results, while KANSVSR further reduces the tOF value. This suggests that the proposed KAM and KFM help to improve cross-frame feature alignment and temporal feature aggregation, leading to more temporally consistent reconstruction results.

Table 3. Temporal consistency comparison among different methods. Lower tOF indicates better temporal consistency.

Method	Bicubic	BasicVSR++	RASVSR	MADNet	KANSVSR
tOF ↓	0.1324	0.0982	0.0974	0.0965	**0.0913**

Note: The arrow after the metric indicates that lower values are better. Bold indicates the best performance.

5.1.3. Qualitative Results

Figures 5 and 6 present qualitative comparisons of different methods on representative satellite video frames. As can be observed, the proposed KANSVSR consistently produces sharper details and more accurate structural reconstructions compared with other approaches.

In Figure 5, which contains fine textual patterns, KANSVSR reconstructs the digits with the highest clarity, while most competing methods generate blurred or distorted results. In particular, BasicVSR exhibits noticeable reconstruction errors, indicating its limitation in handling complex local details. In contrast, the proposed method preserves clear boundaries and sharp structures, demonstrating its superior ability to recover fine-grained information.

In Figure 6, which involves repetitive stripe structures, KANSVSR successfully restores the correct vertical patterns. In comparison, several existing methods produce incorrect diagonal textures or introduce visible artifacts, such as over-smoothed regions or false patterns. These results indicate that KANSVSR is more effective in preserving structural consistency under challenging scenarios.

The superior visual quality of KANSVSR can be attributed to the proposed KAN-based alignment and fusion mechanisms. The KAM enhances the representation of neighboring-frame features, enabling more accurate motion alignment, while the KFM improves the aggregation of temporal information across frames. As a result, the model is able to better exploit temporal cues and recover more faithful spatial structures.

The qualitative results further validate the effectiveness of the proposed method, showing that KANSVSR can generate visually more accurate and structurally consistent reconstructions in complex satellite video scenarios.

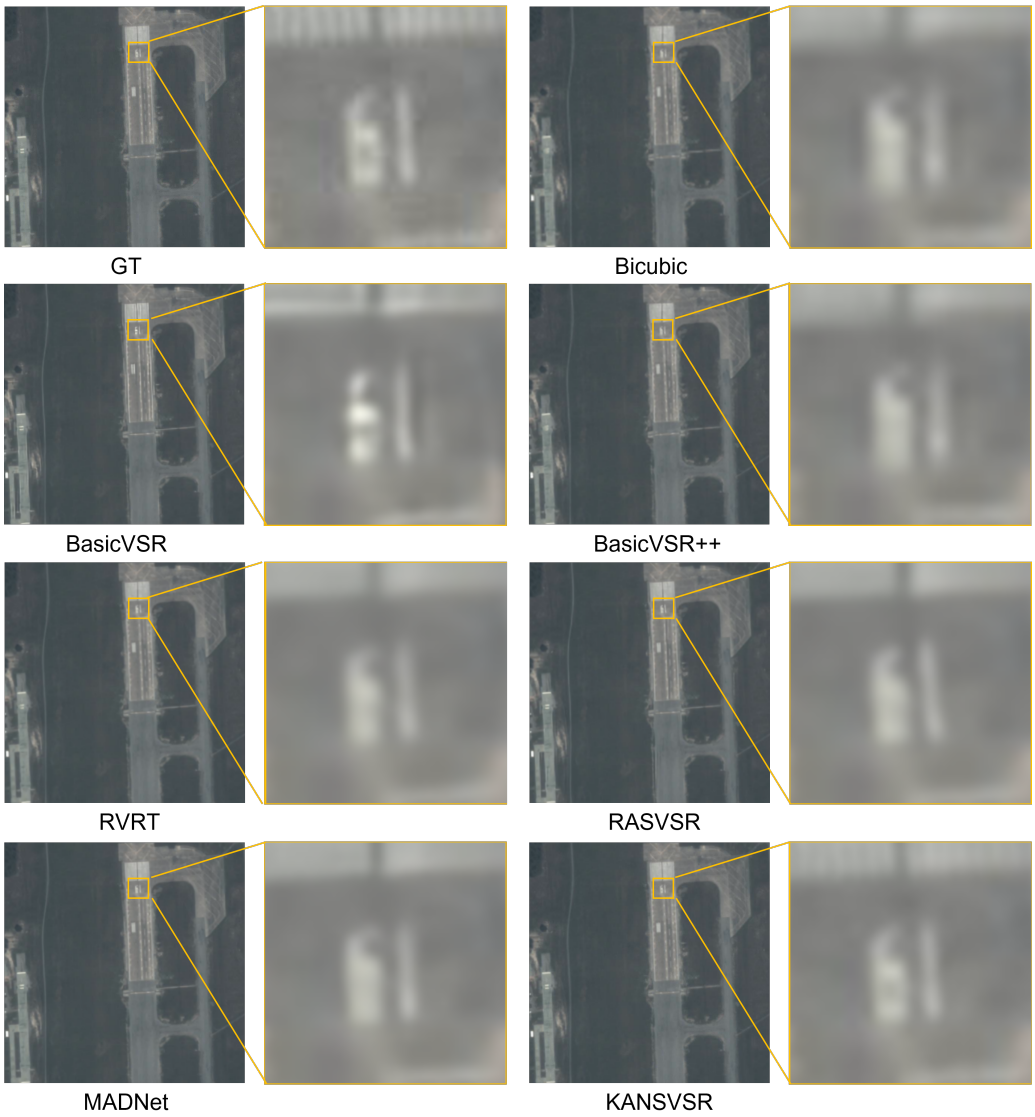

Figure 5. Qualitative comparison. The zoomed-in regions are used to compare the restoration of digit-like details. KANSVSR produces clearer boundaries and more complete local structures, while competing methods show varying degrees of blurring, distortion, or missing details.

5.1.4. Efficiency Analysis

Table 4 and Figure 7 present the comparison of reconstruction performance and computational efficiency among different methods. It can be observed that the proposed KANSVSR achieves superior reconstruction quality while maintaining a relatively lightweight model structure.

Compared with existing methods, KANSVSR achieves the highest PSNR and SSIM while keeping the number of parameters at a relatively low level (5.6 M). Although RASVSR has slightly fewer parameters and lower FLOPs, its reconstruction performance is notably inferior to that of KANSVSR. This indicates that KANSVSR provides a better trade-off between reconstruction accuracy and model complexity.



Figure 6. Qualitative comparison. The zoomed-in regions highlight the preservation of dense vertical patterns. Compared with other methods that introduce over-smoothing, false diagonal textures, or discontinuous stripes, KANSVSR better maintains structural consistency and sharpness.

Table 4. Quantitative comparison of the balance between performance and efficiency among different methods.

Method	PSNR ↑/SSIM ↑	Params (M) ↓	FLOPs (G) ↓
BasicVSR	38.8943/0.9558	6.3	796.6
IconVSR	38.7345/0.9569	8.7	974.7
BasicVSR++	38.8152/0.9580	7.3	880.2
RVRT	38.8078/0.9576	6.2	832.4
RASVSR	39.0090/0.9589	**5.4**	**638.2**
MADNet	**39.1264/0.9607**	6.7	759.9
KANSVSR (Ours)	**39.5676/0.9623**	**5.6**	**665.7**

Note: Metric arrows indicate the preferred direction: ↑ means higher is better, while ↓ means lower is better. The best and second-best results are highlighted in bold red and bold blue, respectively.

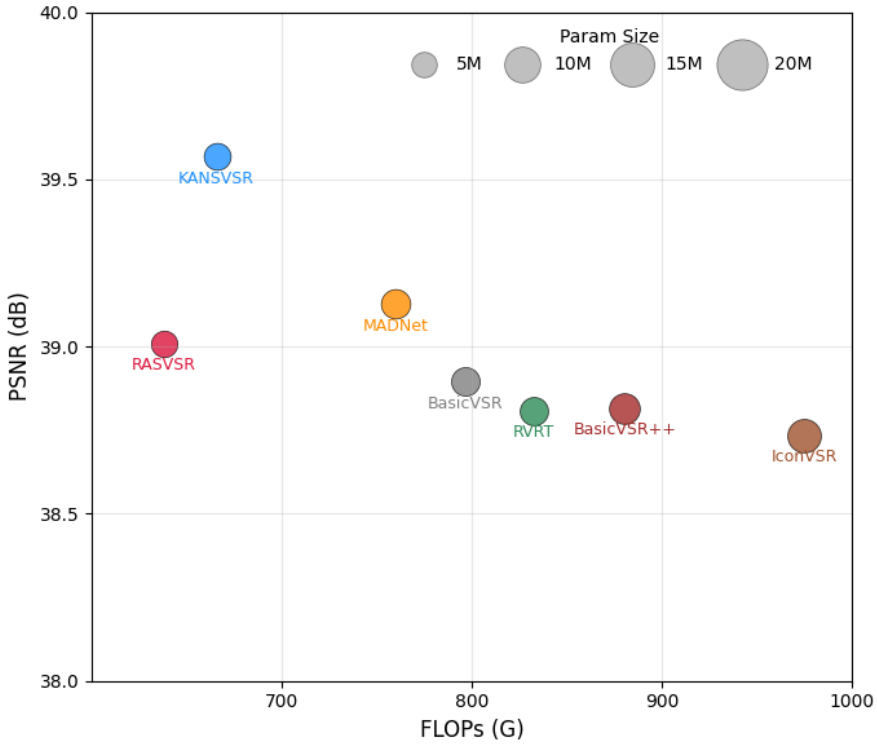

Figure 7. Comparison of FLOPs and PSNR performance across different methods. The *x*-axis denotes FLOPs, the *y*-axis represents PSNR, and the bubble size indicates the parameter count.

In comparison with methods such as BasicVSR, IconVSR, and BasicVSR++, KANSVSR not only achieves higher reconstruction quality but also significantly reduces computational cost. In particular, IconVSR and BasicVSR++ require substantially higher FLOPs while delivering lower PSNR, demonstrating the efficiency advantage of the proposed method.

Furthermore, when compared with MADNet, which is specifically designed for satellite video super-resolution, KANSVSR achieves a higher reconstruction performance with a comparable model size and lower computational cost. This highlights the effectiveness of the proposed framework in balancing performance and efficiency under remote sensing scenarios.

The efficiency of KANSVSR can be mainly attributed to the integration of KAN into the alignment and fusion processes. By enhancing feature representation capability without introducing excessive parameters, the KAN-based modules enable more effective utilization of temporal information. As a result, the model achieves improved reconstruction performance without significantly increasing computational burden.

The proposed KANSVSR demonstrates a favorable balance between accuracy and efficiency, making it well suited for practical satellite video super-resolution applications where both reconstruction quality and computational resources are critical considerations.

5.2. Ablation Study

5.2.1. Ablation Study of KAM and KFM

To evaluate the effectiveness of the proposed KAM and KFM, we conduct ablation experiments by selectively removing these components from the full model. The results are summarized in Table 5.

Table 5. Ablation study of the proposed KAM and KFM modules.

Setting	KAM	KFM	PSNR (dB) ↑	SSIM ↑	LPIPS ↓
w/o KAM & KFM	×	×	38.6724	0.9568	0.0851
w/o KFM	✓	×	39.1835	0.9601	0.0804
w/o KAM	×	✓	39.0572	0.9595	0.0817
Complete Model	✓	✓	**39.5676**	**0.9623**	**0.0777**

Note: ✓ and × indicate use and removal, respectively. Metric arrows indicate the preferred direction: ↑ means higher is better, while ↓ means lower is better. Bold indicates the best performance.

When both KAM and KFM are removed, the network degenerates into a simplified baseline, achieving 38.6724 PSNR and 0.9568 SSIM. Introducing KAM alone leads to a clear performance improvement, reaching 39.1835 PSNR and 0.9601 SSIM. This suggests that enhancing neighboring-frame features through KAN-based spatial attention helps improve alignment accuracy.

When only KFM is retained, the model achieves 39.0572 PSNR and 0.9595 SSIM, indicating that channel-wise feature modulation effectively improves multi-frame feature aggregation. Although the improvement is slightly smaller compared with KAM, it still demonstrates the importance of feature fusion in SVSR.

When both modules are incorporated, the model achieves the best performance, with 39.5676 PSNR, 0.9623 SSIM, and 0.0777 LPIPS. The consistent improvements across all metrics indicate that KAM and KFM are complementary, where KAM focuses on improving alignment quality while KFM enhances feature aggregation. Their combination enables more effective utilization of temporal information, leading to better reconstruction results.

5.2.2. Ablation Study of KAN-Based Feature Modeling

To further analyze the role of KAN in feature modeling during the alignment and fusion stages, we construct several alternative variants under the same overall framework. Specifically, the positions of KAM and KFM, as well as the remaining network architecture, are kept unchanged, while the KAN-related mappings are replaced with MLP, CBAM [68], and self-attention modules, respectively, forming variants with comparable model complexity. All variants are trained and tested under the same settings, and the results are shown in Table 6.

As shown in Table 6, the MLP variant obtains 39.2847 PSNR and 0.9604 SSIM, which are lower than those of the KAN version with 39.5676 PSNR and 0.9623 SSIM. This indicates that standard linear mappings are still limited in modeling complex feature relationships. The CBAM variant enhances features through conventional spatial and channel attention, but its performance remains inferior to that of the KAN version. The self-attention variant achieves relatively better results, but its FLOPs increase to 747.2G, leading to a noticeably higher computational cost than the KAN version.

Table 6. Comparison with alternative feature modeling modules in KAM and KFM.

Variant	Params (M) ↓	FLOPs (G) ↓	PSNR (dB) ↑	SSIM ↑
MLP	5.5	677.2	39.2847	0.9604
CBAM	5.2	665.6	39.2163	0.9600
Self-attention	5.7	747.2	39.3865	0.9610
KAN (Ours)	5.6	665.7	**39.5676**	**0.9623**

Note: Metric arrows indicate the preferred direction: ↑ means higher is better, while ↓ means lower is better. Bold indicates the best reconstruction performance.

Although the alternative variants also benefit from the same overall alignment and fusion framework, the KAN-based variant consistently achieves the best reconstruction

performance. These results show that the flexible feature modeling capability provided by the learnable nonlinear functions in KAN helps improve feature representation in both alignment and fusion stages, thereby further enhancing SVSR reconstruction performance.

5.2.3. Impact of Spline Order

To investigate the impact of spline order, we conduct experiments under different settings while fixing the grid size to 5. The results are presented in Table 7. As the spline order increases, the number of model parameters gradually grows, while the FLOPs remain nearly unchanged. This is because higher-order B-splines introduce additional learnable coefficients without significantly altering the overall computational structure.

Table 7. Quantitative comparison results under different spline orders (grid size fixed to 5).

Spline Order	Params (M) ↓	FLOPs (G) ↓	PSNR (dB) ↑	SSIM ↑
1	5.54	665.65	39.2184	0.9601
3	5.64	665.65	**39.5676**	**0.9623**
5	5.74	665.66	39.5412	0.9621
7	5.83	665.66	39.5087	0.9619
9	5.93	665.66	39.4765	0.9617

Note: Metric arrows indicate the preferred direction: ↑ means higher is better, while ↓ means lower is better. Bold indicates the best reconstruction performance.

In terms of reconstruction performance, the model achieves the best results when the spline order is set to 3, reaching 39.5676 PSNR and 0.9623 SSIM. When the spline order is 1, the limited expressive capacity of low-order B-splines restricts the model's ability to capture complex feature relationships, leading to inferior performance.

As the spline order further increases beyond 3, no additional improvement is observed, and a slight performance degradation occurs. This suggests that increasing the order does not necessarily enhance representation capability for the SVSR task. Instead, higher-order splines may introduce redundant parameters, which can negatively affect model stability and generalization.

These observations indicate that a moderate spline order is sufficient to model the nonlinear feature relationships in SVSR, while avoiding unnecessary model complexity. Therefore, a spline order of 3 is adopted as the default setting in this work.

5.2.4. Impact of Grid Size

To further investigate the effect of grid size, we conduct experiments under different settings while fixing the spline order to 3. The results are summarized in Table 8. As the grid size increases, the number of model parameters grows gradually, while the FLOPs remain nearly unchanged, since the variation mainly affects the number of spline coefficients rather than the overall computational structure.

From the performance perspective, small grid sizes such as 1 or 3 lead to inferior results. This is because a coarse spline grid limits the flexibility of the representation, making it difficult for the model to capture complex feature variations. As the grid size increases, the reconstruction performance improves significantly, and the model achieves the best results when the grid size is set to 5.

Table 8. Quantitative comparison results under different grid sizes (with spline order fixed at 3).

Grid Size	Params (M) ↓	FLOPs (G) ↓	PSNR (dB) ↑	SSIM ↑
1	5.44	665.65	38.2817	0.8724
3	5.54	665.65	39.2946	0.9430
5	5.64	665.65	**39.5676**	**0.9623**
7	5.74	665.66	39.3200	0.9597
9	5.83	665.66	38.7100	0.9310

Note: Metric arrows indicate the preferred direction: ↑ means higher is better, while ↓ means lower is better. Bold indicates the best reconstruction performance.

However, further increasing the grid size does not bring additional benefits and may even result in slight performance degradation. This suggests that overly dense spline grids introduce redundant parameters without effectively improving representation capability, and may even affect model stability.

These results indicate that an appropriate grid density is important for balancing representation capacity and model complexity. A grid size of 5 provides sufficient flexibility for modeling feature variations while avoiding unnecessary parameter growth, and is therefore adopted as the default setting in our model.

6. Conclusions

In this paper, we propose KANSVSR, a lightweight satellite video super-resolution framework tailored for remote sensing scenarios. By incorporating Kolmogorov–Arnold Networks (KAN) into both the alignment and fusion stages, the proposed method enables more effective modeling of cross-frame dependencies.

Specifically, the KAM enhances neighboring-frame representations through spatial attention, leading to more accurate feature alignment under complex motion conditions. Meanwhile, the KFM leverages channel-wise modulation to improve multi-frame feature aggregation, allowing the model to better exploit temporal information. These designs jointly contribute to more precise reconstruction of fine details and more consistent structural recovery.

Extensive experiments demonstrate that KANSVSR achieves superior reconstruction performance in terms of PSNR, SSIM, and LPIPS, while maintaining a compact model size and relatively low computational cost. Compared with existing methods, the proposed framework provides a more favorable balance between accuracy and efficiency, making it suitable for satellite video super-resolution tasks where both performance and resource constraints are critical.

In addition, the experimental analysis verifies the effectiveness of the proposed KAN-based modules and highlights the importance of properly configuring spline order and grid size for balancing representation capacity and model complexity.

For future work, we plan to further explore more efficient KAN-based architectures and extend the proposed framework to handle more realistic and complex degradation models. Investigating its potential in related remote sensing tasks, such as object tracking and scene understanding, is also a promising direction.

Author Contributions: Conceptualization, J.X. and H.Z.; methodology, J.X.; software, J.X.; validation, J.X.; formal analysis, J.X. and H.Z.; investigation, J.X.; resources, H.Z.; data curation, J.X.; writing—original draft preparation, J.X.; writing—review and editing, H.Z. and Z.J.; visualization, J.X.; supervision, H.Z. and Z.J.; project administration, H.Z.; funding acquisition, H.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the National Natural Science Foundation of China (Grant No. 62271017).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The datasets used in this study are publicly available. The SATSOT dataset is available at http://www.csu.cas.cn/gb/kybm/sjlyzx/gcxx_sjj/sjj_wxxl/202106/t20210607_6080256.html (accessed on 1 April 2026), and the SAT-MTB dataset is available at http://www.csu.cas.cn/gb/kybm/sjlyzx/gcxx_sjj/sjj_wxxl/202211/t20221121_6551405.html (accessed on 1 April 2026). The source code of the proposed method will be released at https://github.com/DylanFantast/KANSVSR (accessed on 1 April 2026).

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Feng, J.; Liang, Y.; Zhang, X.; Zhang, J.; Jiao, L. SDANet: Semantic-Embedded Density Adaptive Network for Moving Vehicle Detection in Satellite Videos. *IEEE Trans. Image Process.* **2023**, *32*, 1788–1801. [CrossRef]
2. Xu, C.; Qi, H.; Zheng, Y.; Peng, S. Real-Time Moving Vehicle Detection in Satellite Video Based on Historical Differential Information and Grouping Features. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 1–16. [CrossRef]
3. Carstens, D.; Amer, R. Spatio-temporal analysis of urban changes and surface water quality. *J. Hydrol.* **2019**, *569*, 720–734. [CrossRef]
4. Hamunyela, E.; Hipondoka, M.; Persendt, F.; Sevelia Nghiyalwa, H.; Thomas, C.; Matengu, K. Spatio-temporal characterization of surface water dynamics with Landsat in endorheic Cuvelai-Etosha Basin (1990–2021). *ISPRS J. Photogramm. Remote Sens.* **2022**, *191*, 68–84. [CrossRef]
5. Xiao, C.; An, W.; Zhang, Y.; Su, Z.; Li, M.; Sheng, W.; Pietikäinen, M.; Liu, L. Highly Efficient and Unsupervised Framework for Moving Object Detection in Satellite Videos. *IEEE Trans. Pattern Anal. Mach. Intell.* **2024**, *46*, 11532–11539. [CrossRef]
6. Zhang, H.; Wen, S.; Wei, Z.; Chen, Z. High-Resolution Feature Generator for Small-Ship Detection in Optical Remote Sensing Images. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 1–11. [CrossRef]
7. Li, Y.; Jiao, L.; Huang, Z.; Zhang, X.; Zhang, R.; Song, X.; Tian, C.; Zhang, Z.; Liu, F.; Yang, S.; et al. Deep Learning-Based Object Tracking in Satellite Videos: A comprehensive survey with a new dataset. *IEEE Geosci. Remote Sens. Mag.* **2022**, *10*, 181–212. [CrossRef]
8. Xu, Q.; Zhou, L.; Sheng, W.; Zhang, Q.; An, W. Deep Reinforcement Learning for Multiple Objects Tracking in Satellite Videos. In *Proceedings of the 2025 3rd International Conference on Computer Graphics and Image Processing (CGIP)*; IEEE: Piscataway, NJ, USA, 2025; pp. 56–60. [CrossRef]
9. Li, S.; Sun, X.; Gu, Y.; Lv, Y.; Zhao, M.; Zhou, Z.; Guo, W.; Sun, Y.; Wang, H.; Yang, J. Recent Advances in Intelligent Processing of Satellite Video: Challenges, Methods, and Applications. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2023**, *16*, 6776–6798. [CrossRef]
10. Mojtahedi, F.F.; Yousefpour, N.; Chow, S.H.; Cassidy, M. Deep Learning for Time Series Forecasting: Review and Applications in Geotechnics and Geosciences. *Arch. Comput. Methods Eng.* **2025**, *32*, 3415–3445. [CrossRef]
11. Gopalakrishnan, S.; Choudhury, A. A 'deep' review of video super-resolution. *Signal Process. Image Commun.* **2024**, *129*, 117175. [CrossRef]
12. Xiao, Y.; Yuan, Q.; Jiang, K.; He, J.; Lin, C.W.; Zhang, L. TTST: A Top-k Token Selective Transformer for Remote Sensing Image Super-Resolution. *IEEE Trans. Image Process.* **2024**, *33*, 738–752. [CrossRef] [PubMed]
13. Xiao, Y.; Yuan, Q.; Jiang, K.; Chen, Y.; Zhang, Q.; Lin, C.W. Frequency-Assisted Mamba for Remote Sensing Image Super-Resolution. *IEEE Trans. Multimed.* **2025**, *27*, 1783–1796. [CrossRef]
14. Yao, X.; Zhang, H.; Xie, F.; Cui, Y.; Shi, Z.; Jiang, Z. Activation Map-Based Visual Explanation for Remote Sensing Image Super-Resolution. *IEEE Geosci. Remote Sens. Lett.* **2026**, *23*, 1–5. [CrossRef]
15. Yao, X.; Zhang, H.; Wen, S.; Xie, F.; Cui, Y.; Shi, Z.; Jiang, Z. Uncertainty Estimation Optimization for Reliable Remote Sensing Single-Image Super-Resolution. *IEEE Trans. Geosci. Remote Sens.* **2026**, *64*, 1–16. [CrossRef]
16. Liang, J.; Cao, J.; Fan, Y.; Zhang, K.; Ranjan, R.; Li, Y.; Timofte, R.; Van Gool, L. VRT: A Video Restoration Transformer. *IEEE Trans. Image Process.* **2024**, *33*, 2171–2182. [CrossRef]
17. Shi, S.; Gu, J.; Xie, L.; Wang, X.; Yang, Y.; Dong, C. Rethinking Alignment in Video Super-Resolution Transformers. In *Proceedings of the Advances in Neural Information Processing Systems*; Koyejo, S., Mohamed, S., Agarwal, A., Belgrave, D., Cho, K., Oh, A., Eds.; Curran Associates, Inc.: Red Hook, NY, USA, 2022; Volume 35, pp. 36081–36093.
18. Su, H.; Li, Y.; Xu, Y.; Fu, X.; Liu, S. A review of deep-learning-based super-resolution: From methods to applications. *Pattern Recognit.* **2025**, *157*, 110935. [CrossRef]

19. Xiao, Y.; Yuan, Q.; Jiang, K.; Jin, X.; He, J.; Zhang, L.; Lin, C.W. Local-Global Temporal Difference Learning for Satellite Video Super-Resolution. *IEEE Trans. Circuits Syst. Video Technol.* **2024**, *34*, 2789–2802. [CrossRef]
20. Xu, K.; Yu, Z.; Wang, X.; Mi, M.B.; Yao, A. Enhancing Video Super-Resolution via Implicit Resampling-based Alignment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*; IEEE: Piscataway, NJ, USA, 2024; pp. 2546–2555.
21. Sun, J.; Shen, H.; Yuan, Q.; Zhang, L. Super-Resolution for Remote Sensing Imagery via the Coupling of a Variational Model and Deep Learning. *IEEE Trans. Geosci. Remote Sens.* **2025**, *63*, 1–19. [CrossRef]
22. Yao, X.; Zhang, H.; Wen, S.; Shi, Z.; Jiang, Z. Single-Image Superresolution for RGB Remote Sensing Imagery via Multiscale CNN-Transformer Feature Fusion. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2025**, *18*, 1302–1316. [CrossRef]
23. Wang, Y.; Shao, Z.; Zuo, X.; Lu, T.; Wang, J.; Wang, Y.; Wang, S.; Zhang, Z.; Zhao, X. NSBRNet: Non-Local Spatio-Temporal Bidirectional Recurrent Network for Satellite Video Super-Resolution. *IEEE Trans. Circuits Syst. Video Technol.* **2026**, *36*, 1473–1486. [CrossRef]
24. Kolmogorov, A.N. On the representations of continuous functions of many variables by superposition of continuous functions of one variable and addition. *Dokl. Akad. Nauk USSR* **1957**, *114*, 953–956.
25. Givental, A.B.; Khesin, B.A.; Marsden, J.E.; Varchenko, A.N.; Vassiliev, V.A.; Viro, O.Y.; Zakalyukin, V.M. (Eds.) On the representation of functions of several variables as a superposition of functions of a smaller number of variables. In *Collected Works: Representations of Functions, Celestial Mechanics and KAM Theory, 1957–1965*; Springer: Berlin/Heidelberg, Germany, 2009; pp. 25–46. [CrossRef]
26. Liu, Z.; Wang, Y.; Vaidya, S.; Ruehle, F.; Halverson, J.; Soljačić, M.; Hou, T.Y.; Tegmark, M. KAN: Kolmogorov-Arnold Networks. *arXiv* **2025**, arXiv:2404.19756. [CrossRef]
27. Liu, Z.; Ma, P.; Wang, Y.; Matusik, W.; Tegmark, M. KAN 2.0: Kolmogorov-Arnold Networks Meet Science. *arXiv* **2024**, arXiv:2408.10205. [CrossRef]
28. Somvanshi, S.; Javed, S.A.; Islam, M.M.; Pandit, D.; Das, S. A Survey on Kolmogorov-Arnold Network. *ACM Comput. Surv.* **2025**, *58*, 1–35. [CrossRef]
29. Baniya, A.A.; Lee, T.K.; Eklund, P.W.; Aryal, S. A Survey of Deep Learning Video Super-Resolution. *IEEE Trans. Emerg. Top. Comput. Intell.* **2024**, *8*, 2655–2676. [CrossRef]
30. Schultz, R.; Stevenson, R. Extraction of high-resolution frames from video sequences. *IEEE Trans. Image Process.* **1996**, *5*, 996–1011. [CrossRef]
31. Liu, C.; Sun, D. On Bayesian Adaptive Video Super Resolution. *IEEE Trans. Pattern Anal. Mach. Intell.* **2014**, *36*, 346–360. [CrossRef]
32. Ma, Z.; Liao, R.; Tao, X.; Xu, L.; Jia, J.; Wu, E. Handling Motion Blur in Multi-Frame Super-Resolution. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*; IEEE: Piscataway, NJ, USA, 2015.
33. Dong, C.; Loy, C.C.; He, K.; Tang, X. Image Super-Resolution Using Deep Convolutional Networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2016**, *38*, 295–307. [CrossRef]
34. Jo, Y.; Oh, S.W.; Kang, J.; Kim, S.J. 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 (CVPR)*; IEEE: Piscataway, NJ, USA, 2018.
35. Haris, M.; Shakhnarovich, G.; Ukita, N. Recurrent Back-Projection Network for Video Super-Resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*; IEEE: Piscataway, NJ, USA, 2019.
36. Tian, Y.; Zhang, Y.; Fu, Y.; Xu, C. TDAN: Temporally-Deformable Alignment Network for Video Super-Resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*; IEEE: Piscataway, NJ, USA, 2020.
37. Wang, X.; Chan, K.C.; Yu, K.; Dong, C.; Change Loy, C. EDVR: Video Restoration with Enhanced Deformable Convolutional Networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*; IEEE: Piscataway, NJ, USA, 2019.
38. Chan, K.C.; Wang, X.; Yu, K.; Dong, C.; Loy, C.C. BasicVSR: The Search for Essential Components in Video Super-Resolution and Beyond. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*; IEEE: Piscataway, NJ, USA, 2021; pp. 4947–4956.
39. Chan, K.C.; Zhou, S.; Xu, X.; Loy, C.C. BasicVSR++: Improving Video Super-Resolution with Enhanced Propagation and Alignment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*; IEEE: Piscataway, NJ, USA, 2022; pp. 5972–5981.
40. Liang, J.; Fan, Y.; Xiang, X.; Ranjan, R.; Ilg, E.; Green, S.; Cao, J.; Zhang, K.; Timofte, R.; Gool, L.V. Recurrent Video Restoration Transformer with Guided Deformable Attention. In *Proceedings of the Advances in Neural Information Processing Systems*; Koyejo, S., Mohamed, S., Agarwal, A., Belgrave, D., Cho, K., Oh, A., Eds.; Curran Associates, Inc.: Red Hook, NY, USA, 2022; Volume 35, pp. 378–393.

41. Zhang, S.; Yuan, Q.; Li, J. Video Satellite Imagery Super Resolution for 'Jilin-1' via a Single-and-Multi Frame Ensembled Framework. In *Proceedings of the IGARSS 2020—2020 IEEE International Geoscience and Remote Sensing Symposium*; IEEE: Piscataway, NJ, USA, 2020; pp. 2731–2734. [CrossRef]
42. Xiao, Y.; Su, X.; Yuan, Q.; Liu, D.; Shen, H.; Zhang, L. Satellite Video Super-Resolution via Multiscale Deformable Convolution Alignment and Temporal Grouping Projection. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 1–19. [CrossRef]
43. Wei, X.; Zhang, H.; Jiang, Z. Satellite Video Super-Resolution via Unidirectional Recurrent Network and Various Degradation Modeling. In *Proceedings of the IGARSS 2024—2024 IEEE International Geoscience and Remote Sensing Symposium*; IEEE: Piscataway, NJ, USA, 2024; pp. 6982–6985. [CrossRef]
44. Wang, H.; Li, S.; Zhao, M. A Lightweight Recurrent Aggregation Network for Satellite Video Super-Resolution. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2024**, *17*, 685–695. [CrossRef]
45. Xiao, Y.; Yuan, Q.; Jiang, K.; Chen, Y.; Wang, S.; Lin, C.W. Multi-Axis Feature Diversity Enhancement for Remote Sensing Video Super-Resolution. *IEEE Trans. Image Process.* **2025**, *34*, 1766–1778. [CrossRef]
46. Liu, H.; Gu, Y. Deep Joint Estimation Network for Satellite Video Super-Resolution with Multiple Degradations. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 1–15. [CrossRef]
47. Shen, H.; Qiu, Z.; Yue, L.; Zhang, L. Deep-Learning-Based Super-Resolution of Video Satellite Imagery by the Coupling of Multiframe and Single-Frame Models. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 1–14. [CrossRef]
48. He, Z.; He, D. A Unified Network for Arbitrary Scale Super-Resolution of Video Satellite Images. *IEEE Trans. Geosci. Remote Sens.* **2021**, *59*, 8812–8825. [CrossRef]
49. He, Z.; Li, X.; Qu, R. Video Satellite Imagery Super-Resolution via Model-Based Deep Neural Networks. *Remote Sens.* **2022**, *14*, 749. [CrossRef]
50. He, Z.; He, D.; Li, X.; Xu, J. Unsupervised Video Satellite Super-Resolution by Using Only a Single Video. *IEEE Geosci. Remote Sens. Lett.* **2022**, *19*, 1–5. [CrossRef]
51. Wang, P.; Sertel, E. Multi-frame super-resolution of remote sensing images using attention-based GAN models. *Knowl.-Based Syst.* **2023**, *266*, 110387. [CrossRef]
52. Essahraui, S.; Lamaakal, I.; Makkaoui, K.E.; Ouahbi, I.; Bouami, M.F.; Maleh, Y. Kolmogorov—Arnold Networks: Overview of Architectures and Use Cases. In *Proceedings of the 2025 International Conference on Circuit, Systems and Communication (ICCSC)*; IEEE: Piscataway, NJ, USA, 2025; pp. 1–6. [CrossRef]
53. Vaca-Rubio, C.J.; Blanco, L.; Pereira, R.; Caus, M. Kolmogorov-Arnold Networks (KANs) for Time Series Analysis. In *Proceedings of the 2024 IEEE Globecom Workshops (GC Wkshps)*; IEEE: Piscataway, NJ, USA, 2024; pp. 1–6. [CrossRef]
54. Wang, Y.; Chen, R.; Zhao, Y. KAN-GAN: Awaring Continuity and Frequency for Inertial Signal Super-Resolution. *IFAC-PapersOnLine* **2025**, *59*, 752–757. [CrossRef]
55. Li, C.; Liu, X.; Li, W.; Wang, C.; Liu, H.; Liu, Y.; Chen, Z.; Yuan, Y. U-KAN Makes Strong Backbone for Medical Image Segmentation and Generation. *Proc. AAAI Conf. Artif. Intell.* **2025**, *39*, 4652–4660. [CrossRef]
56. An, Z.; Shi, C.; Wang, Y. Enhanced face super-resolution via iterative collaboration of KAN–MLP hybrid attention and landmark estimation. *Vis. Comput.* **2025**, *42*, 51. [CrossRef]
57. Sabih, M.; Alqarni, A.; Almekkawy, M. KAN-ULM: Advancing Super Resolution Imaging in Ultrasound Localization Microscopy Through Compact Deep Learning Model. In *Proceedings of the 2025 47th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*; IEEE: Piscataway, NJ, USA, 2025; pp. 1–5. [CrossRef]
58. Li, C.; Zhang, B.; Xu, H.; Hong, D. Hyperkan: An Interpretable Approximation Network for Hyperspectral Superresolution. In *Proceedings of the 2024 14th Workshop on Hyperspectral Imaging and Signal Processing: Evolution in Remote Sensing (WHISPERS)*; IEEE: Piscataway, NJ, USA, 2024; pp. 1–5. [CrossRef]
59. Zhao, S.; Bi, M.; Wang, Y.; Teng, X.; Guo, Z.; Chen, K.; Luo, J.; Xu, K.; Hu, M. Hybrid Integration of CNN and Kolmogorov–Arnold Networks for Multispectral Image Super-Resolution. *IEEE Geosci. Remote Sens. Lett.* **2026**, *23*, 1–5. [CrossRef]
60. Ramteke, A.; Arun, P.V.; Srivastava, V.; Buddhiraju, K.M. Low-Rank Spectral–Spatial Super-Resolution of Hyperspectral Images Using KAN-Based GAN. *IEEE Trans. Geosci. Remote Sens.* **2026**, *64*, 1–14. [CrossRef]
61. Zhao, M.; Li, S.; Xuan, S.; Kou, L.; Gong, S.; Zhou, Z. SatSOT: A Benchmark Dataset for Satellite Video Single Object Tracking. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 1–11. [CrossRef]
62. Li, S.; Zhou, Z.; Zhao, M.; Yang, J.; Guo, W.; Lv, Y.; Kou, L.; Wang, H.; Gu, Y. A Multitask Benchmark Dataset for Satellite Video: Object Detection, Tracking, and Segmentation. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 1–21. [CrossRef]
63. Kingma, D.P.; Ba, J. Adam: A Method for Stochastic Optimization. *arXiv* **2017**, arXiv:1412.6980. [CrossRef]
64. Loshchilov, I.; Hutter, F. SGDR: Stochastic Gradient Descent with Warm Restarts. *arXiv* **2017**, arXiv:1608.03983. [CrossRef]
65. Charbonnier, P.; Blanc-Feraud, L.; Aubert, G.; Barlaud, M. Two deterministic half-quadratic regularization algorithms for computed imaging. In *Proceedings of the 1st International Conference on Image Processing*; IEEE: Piscataway, NJ, USA, 1994; Volume 2, pp. 168–172. [CrossRef]

66. Zhang, R.; Isola, P.; Efros, A.A.; Shechtman, E.; Wang, O. The Unreasonable Effectiveness of Deep Features as a Perceptual Metric. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*; IEEE: Piscataway, NJ, USA, 2018.
67. Chu, M.; Xie, Y.; Mayer, J.; Leal-Taixé, L.; Thuerey, N. Learning temporal coherence via self-supervision for GAN-based video generation. *ACM Trans. Graph.* **2020**, *39*, 75:1–75:13. [CrossRef]
68. Woo, S.; Park, J.; Lee, J.Y.; Kweon, I.S. CBAM: Convolutional Block Attention Module. In *Proceedings of the European Conference on Computer Vision (ECCV)*; Springer: Berlin/Heidelberg, Germany, 2018.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.