Hazy Low-Quality Satellite Video Restoration Via Learning Optimal Joint Degradation Patterns and Continuous-Scale Super-Resolution Reconstruction

Ning Ni Libao Zhang*
School of Artificial Intelligence, Beijing Normal University, Beijing 100875, China
ningni@mail.bnu.edu.cn libaozhang@bnu.edu.cn

Abstract

Currently, the demand for higher video quality has grown significantly. However, satellite video has low resolution, complex motion, and weak textures. Haze interference further exacerbates the loss of motion information and texture details, hindering effective spatiotemporal feature fusion and fine-grained feature mining. This presents significant challenges for subsequent super-resolution (SR) reconstruction, especially at continuous scales. To address these problems, this paper models the double-degradation process of hazy low-quality satellite videos and proposes a novel network to learn the optimal joint degradation pattern (ODPNet) for continuous-scale SR of hazy satellite videos. First, we design a prior-based feature soft dehazing module to eliminate haze interference at the feature level. Second, we develop a spatiotemporal self-attention (SSA) to capture long-range feature dependencies, thereby achieving effective spatiotemporal feature fusion. Third, we devise a tri-branch cross-aggregation block (TCB) to enhance feature representations of weak textures in satellite videos by effectively aggregating contextual information. Finally, we propose a cross-scale feature Top-k selection Transformer (CFTST), which aims to adaptively select and aggregate cross-scale latent codes to learn feature representations of satellite videos at arbitrary resolutions, thus enabling continuous-scale SR. Experiments show that ODPNet outperforms existing methods and achieves a better balance between model parameters and performance.

1. Introduction

With the rapid development of remote sensing data processing technology, the demand for high-quality satellite video has grown increasingly urgent. However, on the one hand, the spatial resolution of satellite video is inherently low due to the influence of the data acquisition and transmission process; on the other hand, satellite video imaging is easily affected by haze, resulting in blurred scenes and loss of details. To overcome these issues, dehazing and video super-resolution (VSR) techniques are usually used to improve video quality, thereby providing high-quality data for subsequent tasks or applications such as target tracking [27, 49], object segmentation [15, 47], and moving vehicle detection [24, 43]. Existing satellite VSR methods are designed for single-degraded factors and can achieve admirable results when applied to specific single-degraded conditions. When faced with hazy low-resolution (LR) satellite videos with complex double-degradation, existing VSR methods [2, 21, 31, 35, 48] cannot effectively achieve dehazing and VSR at the same time, and lack the ability to solve the double-degradation in a unified framework.

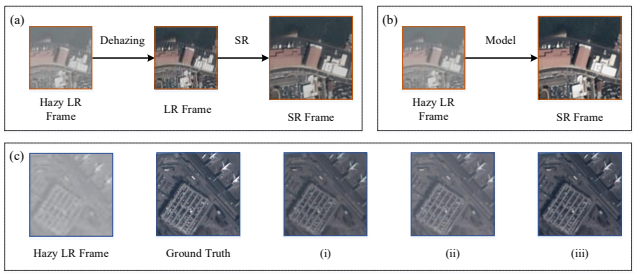

Figure 1. (a) Description of the two-stage process. (b) Description of the end-to-end process. (c) Sample results of our method compared with others. (i) FFANet + LGTD [25, 41]. (ii) DehazeFormer + MSDTGP [30, 40]. (iii) ODPNet (Ours).

To achieve hazy satellite VSR, a simple idea is to separate dehazing and VSR into two independent tasks and then cascade their methods and process them in stages, which is called a two-stage process (TSP). The specific process of TSP is shown in Fig. 1. However, TSP methods cannot directly learn the mapping relationship between double-degraded data and the corresponding high-quality label, which easily leads to the continuous accumulation of estimation errors in each stage. Specifically, satellite video have weak textures, and the haze will further weaken the details, making it difficult to capture and fuse inter-frame

*Corresponding Author. This work is supported by the National Natural Science Foundation of China under Grant 62271060 and 61571050.

information and mine detailed features. In addition, in the practical application of satellite video, it is common to continuously zoomed-in views. However, most existing satellite VSR methods [20, 40–42] restrict the upsampling factor to integers (*e.g.*, $4\times$), which greatly limits the potential applications and deployment options of the model.

To solve the above problems, we model and analyze the double-degradation process of hazy LR satellite video and propose an end-to-end network to learn the optimal joint degradation pattern (ODPNet) for hazy satellite video continuous-scale SR. Specifically, first, we design a prior-based feature soft dehazing module (PFSDM), which uses the imaging mechanism of haze to reversely decode features, thereby eliminating the haze interference at the feature level. Second, we develop a novel spatiotemporal self-attention (SSA) by integrating spatial and temporal self-attention in a novel way, which can effectively capture the long-range dependencies of spatiotemporal features of satellite videos and complete the effective fusion of inter-frame complementary information. Third, we devise a tri-branch cross-aggregation block (TCB) to comprehensively mine detailed information by leveraging interactive spatial and channel self-attention, enabling deep feature aggregation and enhancing feature representation. Finally, we propose a cross-scale feature Top-k selection Transformer (CFTST), which employs a Top-k selection mechanism to adaptively choose the most relevant top k latent codes among multiple scale features to complete the mapping from multi-scale space to continuous-scale space. In addition, we propose a joint optimization strategy to enhance the stability of model training.

The main contributions of this paper are summarized as:

1) We design a novel end-to-end network to handle double-degraded problem. Our ODPNet models the double-degradation process and learns an optimal solution to achieve continuous-scale SR of hazy satellite videos.

2) We propose a new prior-based feature soft dehazing module (PFSDM). Based on the imaging mechanism of haze, our PFSDM effectively eliminates haze interference at the feature level through inverse feature decoding.

3) We develop a spatiotemporal self-attention (SSA) and a tri-branch cross-aggregation block (TCB). SSA achieves a weighted and effective fusion of spatiotemporal complementary information. TCB can enrich feature details and achieve deep feature aggregation.

4) We propose a cross-scale feature Top-k selection Transformer (CFTST). Our CFTST can adaptively select the most relevant top k feature latent codes to complete the continuous-scale mapping.

2. Related Works

Remote sensing image dehazing: Currently, the mainstream remote sensing image (RSI) dehazing methods can be divided into prior-based methods [1, 8, 17, 28] and convolutional neural network (CNN)-based methods [3, 5–7]. The prior-based methods usually use dark channel prior [9] or filters to generate haze-free RSI. Singh *et al*. [29] proposed an evolutionary visibility restoration model for RSI, which uses dynamic differential evolution to optimize the parameters in the ASM. Liang *et al*. [18] develop an atmospheric light value-map fusion estimation model that integrates the heterogeneous priors-guided haze concentration model to derive the global atmospheric light value. The CNN-based methods directly use the network to learn the mapping relationship between hazy RSI and haze-free RSI. Zheng *et al*. [46] proposed an enhanced attention-guide generative adversarial network, which used unpaired training data to optimize. Zhang *et al*. [44] proposed the proxy and cross-stripes integration Transformer for RSI dehazing, which achieves a large receptive field with cost-effective computational complexity. Although the above dehazing methods can achieve good performance, they are proposed for single-degraded conditions, when faced with double-degraded data, the performance cannot be well utilized.

Satellite VSR: Liu *et al*. [20] constructed the first effective satellite VSR framework based on local spatiotemporal neighbors and nonlocal similarity modeling. Then, to improve the fusion effect of spatio-temporal information, Xiao *et al*. [40] proposed a novel fusion strategy of temporal grouping projection with deformable convolution. In the same year, Liu *et al*. [19] proposed a deep joint estimation network to deal with the problem that the blur kernel is usually unknown. After that, to improve the ability to model long-distance correspondence, Jin *et al*. [13] proposed a novel two-branch alignment network with an efficient fusion module for satellite VSR. Recently, Xiao *et al*. [41] proposed an efficient time compensation method based on short-term and long-term time difference modeling, which significantly improved the performance of satellite video SR. Although satellite VSR researchers have made significant progress, they cannot simultaneously capture the long-distance dependency of information in the spatiotemporal dimension. In this paper, inspired by [32], we proposed SSA, which can effectively capture the long-distance dependency of spatiotemporal information and complete the effective fusion of inter-frame features.

Continuous-scale SR: Most existing satellite VSR methods limit the upsampling scales to specific integer values and are required to train a distinct model for each upsampling scale. To solve this limitation, Ni *et al*. [23] proposed a novel network for continuous-scale satellite VSR, which achieved good performance on both fixed-scale factors and continuous-scale factors. Although there are few studies on continuous-scale VSR of satellite videos, many plug-and-play continuous-scale upsampling modules [4, 12, 16, 38] have been proposed in the field of single-image SR, such

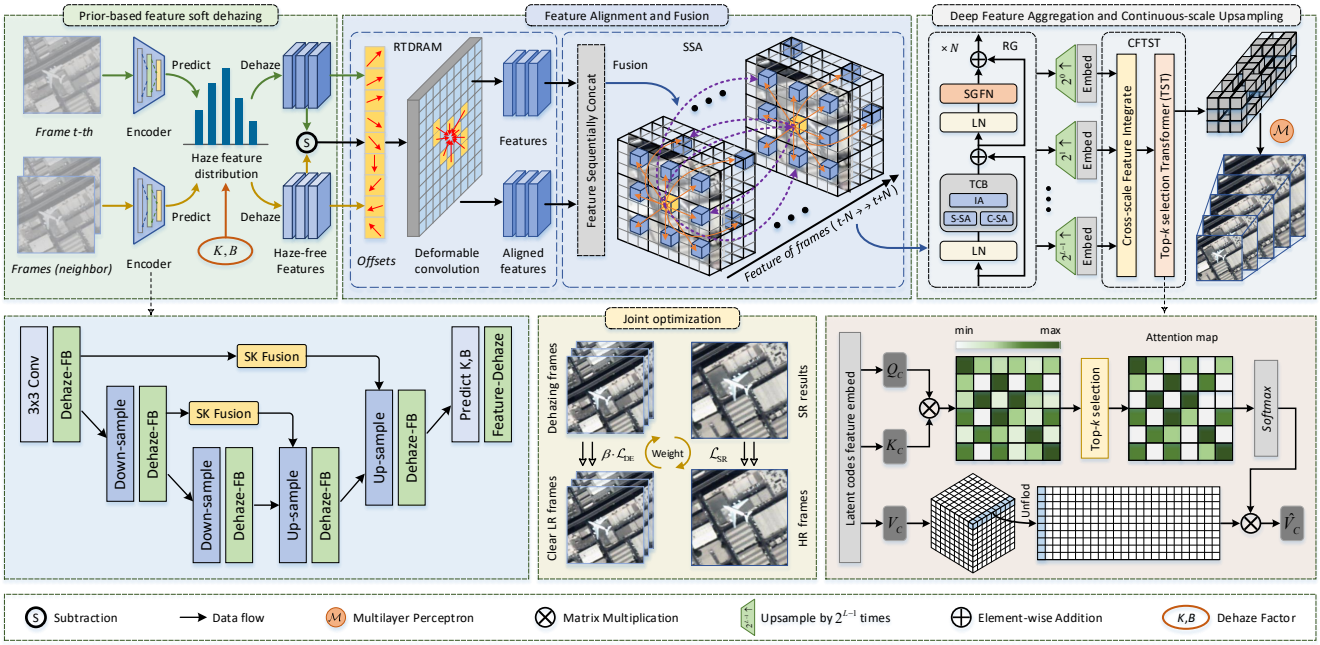

Figure 2. The overall architecture of the proposed ODPNet. Residual group (RG).

as LIFF [4], SADN [38]. After that, Wei *et al*. [37] introduced the concept of neural operators for continuous-scale SR. Recently, to improve computational efficiency, many lightweight continuous-scale SR methods [10, 11, 39] were proposed. Although the above methods have achieved good performance, they integrate features based on coordinate distance when performing feature mapping, and fail to adaptively select the most relevant latent codes (nearest latent codes are not necessarily relevant), resulting in artifacts of the SR results.

3. Methodology

3.1. Degradation Analysis of Hazy Satellite VSR

As we all know, in the typical VSR framework, the degradation process of satellite videos is represented as:

$$LR = (HR \otimes k)\downarrow_s + n, \tag{1}$$

where LR is LR videos, HR is HR videos, k is a blurry kernel, $\otimes$ is a convolution operation, $\downarrow_s$ is the downsampling operator with scale factor s, and n is the independent noise. Unlike the typical VSR degradation process, the haze degradation process is usually simulated using the atmospheric scattering model [22], which can be depicted as follows:

$$I(x) = J(x)\,t(x) + A\,(1 - t(x)), \tag{2}$$

where x represents the position of the pixel, I is the hazy image, J is the clear image (restored image), A is the atmospheric light, t is the transmission rate, which represents the attenuation coefficient of light passing through the haze.

Due to the fundamental differences between Equation (1) and Equation (2), they are basically independent. Therefore, in the framework of hazy satellite VSR, the degradation process can be regarded as their combination, which can be defined as:

$$I_{\mathrm{LR}}(x) = f(LR, I(x)), \tag{3}$$

where $f(\cdot,\cdot)$ denotes the combination function, I_{LR} is the LR hazy video. In the real-world acquisition of satellite videos, haze degradation usually occurs first, the subsequent degradation involves sensor noise and loss of resolution. Therefore, the Equation (3) can be redefined as:

$$I_{\mathrm{LR}}(x) = [(J_{\mathrm{HR}}(x) \cdot t(x) + A \cdot (1 - t(x))) \otimes k]\downarrow_s + n, \tag{4}$$

where J_{HR} denotes the clear satellite video (restored satellite video). Based on the above analysis, the task of hazy satellite VSR needs to address the combined degradation problem of satellite videos. Therefore, designing a model for this task is more complex and challenging.

3.2. Overview

As shown in Fig. 2, our ODPNet aims to use an end-to-end network to achieve dehazing and continuous-scale SR for satellite videos. Due to haze interference, it becomes challenging for the network to capture inter-frame complementary information. Therefore, we propose to remove the haze interference at the feature level before performing inter-frame feature alignment and fusion. Specifically, given a group of hazy LR video frames, including $2N$ supporting frames $\left\{I_{\mathrm{LR}}^{t-N}, \cdots, I_{\mathrm{LR}}^{t-1}, I_{\mathrm{LR}}^{t+1}, \cdots, I_{\mathrm{LR}}^{t+N}\right\}$ and a

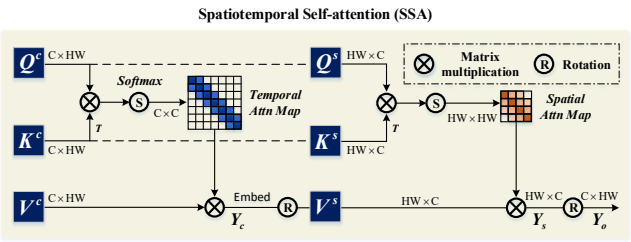

Figure 3. Illustration of the SSA. T is the matrix transpose.

reference frame I_{LR}^{t}. First, in the prior-based feature soft dehazing part, we use an encoder to learn the feature distribution of haze and predict a set of dehazing factors K and B, achieving feature-level dehazing and obtaining a group of haze-free features $F_{\mathrm{LR}}^{[t-N:t+N]}$.

Second, in the feature alignment and fusion part, we introduce the RTDRAM [23] to perform feature alignment due to its strong superiority and obtain the aligned features $F_{\mathrm{aligned}}^{[t-N:t+N]}$. Then, we sequentially concatenate the aligned features along the channel dimension and send them into SSA to achieve effective feature fusion.

Finally, in the deep feature aggregation and continuous-scale upsampling part, the TCB performs deep feature aggregation to further enrich texture details and obtain final features F_{out}^{t}. After that, we send F_{out}^{t} into CFTST to obtain the scale-arbitrary SR results.

3.3. Prior-based Feature Soft Dehazing

As we all know, the degradation of haze is usually described by the atmospheric scattering model [22], which can be depicted as Equation (2). Based on this prior knowledge, if we want to obtain the restored image J, we can rearrange the Equation (2) as follows:

$$J\left(x\right)=\left(\frac{1}{t\left(x\right)}-1\right)I\left(x\right)+\left(1-\frac{1}{t\left(x\right)}\right)A+I\left(x\right), \quad (5)$$

let $K\left(x\right)=\frac{1}{t(x)}-1$, $B\left(x\right)=\left(1-\frac{1}{t(x)}\right)A$, the Equation (5) can be rewritten as:

$$J\left(x\right)=K\left(x\right)I\left(x\right)+B\left(x\right)+I\left(x\right), \quad (6)$$

therefore, the network only needs to predict $K\left(x\right)$ and $B\left(x\right)$, and a clear restored image can be obtained through the Equation (6). It should be noted that we need to obtain the haze-free feature in the ODPNet, so we only need to replace $I\left(x\right)$ in Equation (6) with the shallow feature $\bar{I}\left(x\right)$ of the $I\left(x\right)$. Therefore, at the feature level, the Equation (6) can be rewritten as:

$$\bar{J}\left(x\right)=K\left(x\right)\bar{I}\left(x\right)+B\left(x\right)+\bar{I}\left(x\right), \quad (7)$$

where $\bar{J}\left(x\right)$ is the haze-free feature of the $I\left(x\right)$. Here, the satellite video frames $I_{\mathrm{LR}}^{[t-N:t+N]}$ are sent into encoder to predict K and B for feature-level dehazeing, as follows:

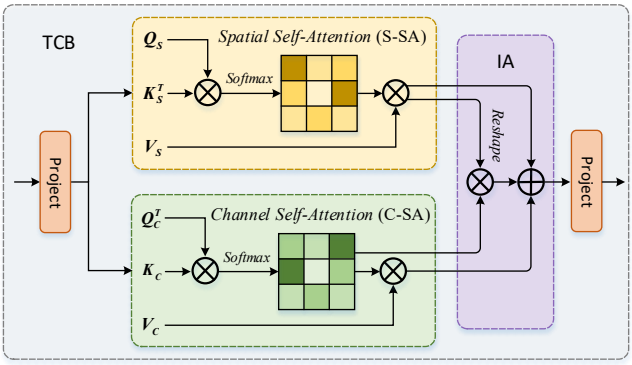

Figure 4. The detailed structure of the TCB.

$$K,B=f_{\mathrm{Enocder}}\left(I_{\mathrm{LR}}^{[t-N:t+N]}\right), \quad (8)$$

$$F_{\mathrm{LR}}^{[t-N:t+N]}=K\cdot F_{\mathrm{Haze}}^{[t-N:t+N]}+B+F_{\mathrm{Haze}}^{[t-N:t+N]}, \quad (9)$$

where f_{Enocder} denotes the prediction function of the encoder, $F_{\mathrm{Haze}}^{[t-N:t+N]}$ is the shallow features of $I_{\mathrm{LR}}^{[t-N:t+N]}$, which can be extracted by a convolutional layer. $F_{\mathrm{LR}}^{[t-N:t+N]}$ is the haze-free features. The encoder adopts a U-shaped structure, and the details are shown in Fig. 2. Since the effectiveness of dehazeformer block (Dehaze-FB) has been proven [30], we introduce it into our encoder. In addition, we apply the SK fusion [30] module to fuse features of different scales generated by Dehaze-FBs.

3.4. Spatiotemporal Self-attention

Satellite video textures are weak, and feature-level dehazing will further weaken the feature representation. Therefore, we design an SSA to achieve effective feature fusion by capturing long-range feature dependencies. The detailed structure of the SSA is shown in Fig. 3.

Specifically, we first concatenate the aligned features $F_{\mathrm{aligned}}^{[t-N:t+N]}$ in the channel dimension according to temporal order, thereby incorporating temporal information in the channel dimension. Suppose $\mathcal{Y}\in\mathbb{R}^{H\times W\times C}$ denotes the input feature, where H and W are the height and width of the feature respectively, and C is the number of channels. First, $\mathcal{Y}$ are reshaped and projected into the query matrixs $Q\in\mathbb{R}^{C\times HW}$, key matrixs $K\in\mathbb{R}^{C\times HW}$, and value matrixs $V\in\mathbb{R}^{C\times HW}$. Then, we perform matrix multiplication of the query matrix Q and key matrix K to obtain the temporal attention map of size $\mathbb{R}^{C\times C}$ to capture inter-frame long-range dependencies. After that, temporal self-attention is applied to obtain the initial fusion result Y_c:

$$Q^{c}=\mathcal{Y}\cdot\omega_{q},K^{c}=\mathcal{Y}\cdot\omega_{k},V^{c}=\mathcal{Y}\cdot\omega_{v}, \quad (10)$$

$$Y_{c}=\xi\left(Q^{c}\otimes\left(K^{c}\right)^{T}\right)\otimes V^{c}, \quad (11)$$

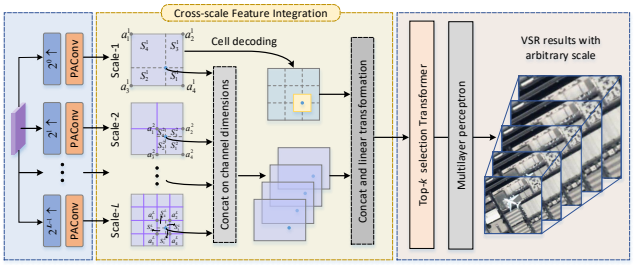

Figure 5. The detailed structure of our proposed CFTST.

where $\omega_q, \omega_k, \omega_v$ denotes the linear projection matrices of query, key, and value, respectively, and $\otimes$ is the matrix multiplication operation. ξ denotes the operation of softmax. Finally, we perform feature embedding and rotation on Y_c to obtain query, key, and value matrices $Q^s, K^s, V^s \in \mathbb{R}^{HW \times C}$ of spatial self-attention. The spatial attention map of size $\mathbb{R}^{HW \times HW}$ is obtained by calculating query matrix Q^s and key matrix K^s to model the intra-frame long-range dependencies. By inversely rotating the spatial attention output Y_s, we get the final fusion result Y_o, as follows:

$$Q^s, K^s, V^s = \mathcal{R}\left(\varphi\left(Y_c\right)\right), \tag{12}$$

$$Y_s = \xi\left(Q^s \otimes \left(K^s\right)^T\right) \otimes V^s, \tag{13}$$

$$Y_o = \mathcal{R}^{-1}\left(Y_s\right), \tag{14}$$

where $\mathcal{R}\left(\cdot\right)$ is the rotation operation, $\mathcal{R}^{-1}\left(\cdot\right)$ is the inverse rotation operation, and φ is feature embedding function.

3.5. Tri-branch Cross-aggregation Block

As shown in Fig. 4, our TCB includes three branches: the spatial self-attention branch, the channel self-attention branch, and their interaction aggregation (IA) branch.

Given the input $X \in \mathbb{R}^{H \times W \times C}$, the query matrix Q, key matrix K and value matrix V are first generated by linear projection. In the spatial self-attention branch, the size of each matrix is in the form of $\mathbb{R}^{HW \times C}$, and each matrix is split into small windows containing N_w pixels and reshaped into Q_S, K_S, and V_S. The same strategy is adopted in the channel self-attention branch, we can obtain Q_C, K_C, and V_C. The calculations of the two branches are as follows:

$$Y_S = \xi\left(Q_S \otimes \left(K_S\right)^T + D\right) \otimes V_S, \tag{15}$$

$$Y_C = V_C \otimes \xi\left(\left(Q_C\right)^T \otimes K_C / \alpha\right), \tag{16}$$

where D represents relative position encoding [34], α is a temperature parameter. Finally, the Y_S and channel attention map are sent into the IA branch to achieve deep feature interaction and aggregation, as follows:

$$Y_{S,C} = \text{Reshap}\left(Y_S\right) \otimes \xi\left(\left(Q_C\right)^T \otimes K_C / \alpha\right), \tag{17}$$

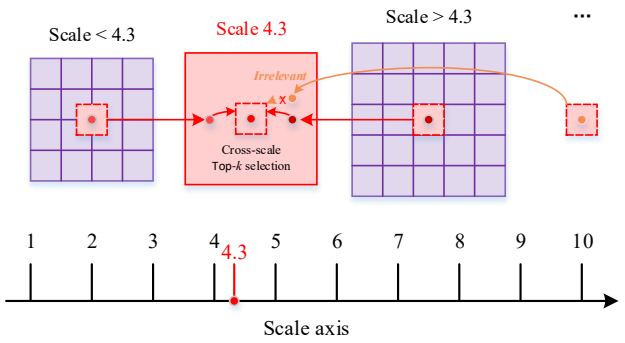

Figure 6. Illustration of cross-scale feature Top-k selection.

$$F_{\text{out}}^t = Y_S + Y_C + Y_{S,C}, \tag{18}$$

where Reshap $\left(\cdot\right)$ represents the reshape operation.

3.6. Cross-scale Feature Top-k Selection Transformer

As shown in Fig. 5, we first perform upsampling and use PAConv [45] to generate a group of cross-scale feature maps $\{F_{\text{map}}^l \in \mathbb{R}^{C \times H(2^{l-1}) \times W(2^{l-1})}\}_{l=1}^L$. Then, we perform cross-scale feature integration. Specifically, given the upsampling scale factor is r, we generate the corresponding 2D coordinates $\mathcal{Z}_r$. Taking an arbitrary query coordinate $z \in \mathcal{Z}_r$, it has the nearest four latent codes at each scale. Taking scale-l as an example, we denote the nearest four latent codes as $\{a_i^l\}_{i=1}^4$. To obtain the feature representation of z, we propose to weight the four latent codes by using the corresponding interpolation factors s_i^l, as follows:

$$a_z^l = \left\{s_i^l \cdot a_i^l\right\}_{i=1}^4, \tag{19}$$

where a_z^l denotes the feature representation (latent code) of coordinate z. The size of s_i^l is related to the area enclosed by the query coordinates z and the nearest coordinates of $\{a_i^l\}_{i=1}^4$, the larger the area, the smaller the weighted interpolation factors s_i^l. By performing Equation (19) on all scales, we can obtain a group of feature representations $\{a_z^l\}_{l=1}^L$ at query coordinate z. Then, we concatenate $\{a_z^l\}_{l=1}^L$ along the channel dimension. After that, inspired by [33], we use Top-k selection Transformer (TST) to perform feature consistency checks and select k most relevant latent codes in the channel dimension, thereby filtering out irrelevant or noisy features, and the idea is shown in Fig. 6.

The structure of TST is shown in Fig. 2, we first embed the input features to obtain the query matrix Q_C, key matrix K_C, and value matrix V_C. Then, we use Q_C and K_C to calculate the relevance of latent codes on the channel dimension to obtain attention map M of size $\mathbb{R}^{C \times C}$. Then, we mask the irrelevant elements assigned with lower attention weights in M. The calculation of M is as follows:

$$M = \left(\left(Q_C\right)^T \otimes K_C\right) / \alpha. \tag{20}$$

Clips	Ave (A, t)	Methods	#Params	×2.1	×2.4	×2.7	×3.2	×3.6	×3.9	×8
Clip001	(0.7, 0.35) (0.8, 0.45) (0.9, 0.55) (1.0, 0.60)	TSP-FTM	4.5+2.4M	29.45/0.9574	27.91/0.9440	26.09/0.9221	23.85/0.9006	21.64/0.8387	21.21/0.8228	16.77/0.5040
		TSP-DEL	4.6+20.2M	31.72/0.9728	29.91/0.9617	27.29/0.9391	25.68/0.9258	22.67/0.8629	21.72/0.8518	21.62/0.6559
		TSP-MEM	7.4+20.6M	29.00/0.9751	29.47/0.9647	27.75/0.9413	27.26/0.9274	24.20/0.8633	24.00/0.8534	20.37/0.6456
		TSP-FC	4.5+10.6M	29.47/0.9580	27.92/0.9448	26.08/0.9234	23.85/0.9043	21.64/0.8437	21.22/0.8293	16.68/0.5242
		TSP-DC	4.6+10.6M	31.69/0.9727	29.92/0.9618	27.32/0.9399	25.73/0.9286	22.70/0.8668	21.75/0.8557	21.73/0.6726
		TSP-MC	7.4+10.6M	29.01/0.9752	29.48/0.9649	27.75/0.9418	27.38/0.9305	24.28/0.8690	24.10/0.8590	20.64/0.6841
		ODPNet*	7.3M	33.65/0.9916	32.48/0.9889	31.64/0.9849	31.84/0.9767	29.18/0.9627	29.14/0.9488	22.27/0.7013
Clip002	(0.7, 0.55) (0.8, 0.60) (0.9, 0.45) (1.0, 0.35)	TSP-FEL	4.5+20.2M	29.56/0.9823	27.45/0.9736	25.58/0.9568	24.73/0.9456	23.90/0.9075	23.81/0.8998	21.67/0.7792
		TSP-DTM	4.6+2.4M	31.63/0.9838	28.89/0.9756	28.47/0.9608	26.83/0.9545	24.82/0.9096	24.99/0.9019	23.36/0.7769
		TSP-MTL	7.4+2.3M	32.96/0.9831	31.78/0.9756	29.73/0.9618	29.60/0.9527	26.42/0.9099	25.93/0.9020	23.73/0.7987
		TSP-FC	4.5+10.6M	29.55/0.9823	27.44/0.9737	25.57/0.9572	24.74/0.9472	23.92/0.9103	23.83/0.9034	21.77/0.7837
		TSP-DC	4.6+10.6M	31.60/0.9843	28.87/0.9761	28.46/0.9620	26.86/0.9584	24.88/0.9179	25.15/0.9182	24.12/0.8328
		TSP-MC	7.4+10.6M	33.00/0.9834	31.77/0.9759	29.74/0.9631	29.77/0.9572	26.55/0.9198	26.13/0.9170	24.36/0.8354
		ODPNet*	7.3M	38.39/0.9923	37.56/0.9906	36.19/0.9881	35.88/0.9808	33.54/0.9701	32.72/0.9624	26.26/0.8302

Table 1. Quantitative comparison (PSNR/SSIM) for continuous-scale satellite VSR with TSP-based methods on two video clips. Red and blue text indicates the best and second best performance respectively. Ave denotes the average result over all (A, t). * indicates our method.

After that, an adaptive Top-k selection strategy is implemented upon M. Here, k is an adjustable parameter to control the number of latent codes in the channel dimension that participate in the final pixel value prediction. Take $k = 0.75$ as an example, only the elements with top 75% attention weights of the M can participate in subsequent weighted calculations. For other remaining elements that are smaller than the top 75% attention weights, we replace the attention weights with 0, as follows:

$$\hat{w}_{ij} = \begin{cases} 1 \cdot w_{ij}, & w_{ij} \in [M]^{k}_{\text{Top}} \\ 0 \cdot w_{ij}, & \text{otherwise} \end{cases}, \forall w_{ij} \in M, \tag{21}$$

where $[M]^{k}_{\text{Top}}$ denotes the Top-k selection of elements in M, $\hat{w}_{ij} \in \hat{M}$, $\hat{M}$ is a new attention map after performing Top-k selection on M. Considering that a single Top-k selection may miss information, to improve the robustness of Top-k, we select a group of Top-k to prevent information loss. Given a group of threshold value $\{k_i\}^{n}_{i=1}$ of Top-k, we perform Equation (21) to obtain a group of attention maps $\{\hat{M}_i\}^{n}_{i=1}$. Finally, we achieve feature weighting by:

$$\hat{V}_C = \sum_{i=1}^{n} \left(V_C \otimes \hat{M}_i \right), \tag{22}$$

where $\hat{V}_C$ is the most relevant feature latent codes. Finally, we use a multi-layer perceptron to complete the mapping of features to pixel values and obtain the SR result $I^{(r)}_{\text{SR}}$.

3.7. Joint Optimization and Loss Functions

To ensure the effectiveness of feature soft dehazing, we propose a loss function that contains joint optimization of dehazing supervision and VSR supervision, as follows:

$$\mathcal{L} = \mathcal{L}_{\text{SR}} + \beta \cdot \mathcal{L}_{\text{DE}}, \tag{23}$$

where $\beta = 0.1$ is the hyperparameter to balance the two sub-items. $\mathcal{L}_{\text{SR}} = \boldsymbol{E}_{r \sim \mathcal{U}(1,4)} \left\| \mathcal{PS}\left(I^{t}_{\text{SR}}\right)^{m}_{r} - \mathcal{PS}\left(I^{t}_{\text{HR}}\right)^{m}_{r} \right\|_1$ and $\mathcal{L}_{\text{DE}} = \sum_{i=t-N}^{t+N} \left\| I^{i}_{\text{LR}} - f_\theta(F^{i}_{\text{LR}}) \right\|_1$. $\mathcal{U}\left(1, 4\right)$ denote a uniform distribution on $[1, 4]$, $\mathcal{PS}\left(\cdot\right)^{m}_{r}$ represents the pixel sampling operation under the scale factor r, and the number of pixel samples is m. f_θ is a function that maps haze-free features to pixel values.

4. Experiments

4.1. Experimental Setup

Datasets and Metrics. Our video data comes from the Jilin-1 satellite and contains nine video scenes, with a resolution of approximately 1 m. Six scenes are used to create the training set, while three scenes are designated for the test set. In constructing the training dataset, we only use the first 110 frames from long-duration video scenes and the first 60 frames from short-duration video scenes. Specifically, the training dataset is cropped from the six video scenes with a size of 640×640. For the test set, we crop from the remaining video scenes to obtain ten video clips (Clip001-Clip010). Furthermore, to objectively demonstrate the performance of our method, we introduce two evaluation metrics: peak signal-to-noise ratio (PSNR) and structural similarity index measure (SSIM) [36].

Implementation Details. The channel numbers of DehazeFBs are set to $[24, 48, 96, 48, 24]$. SSA number is set to 3 and its channel number is 160. The number of RG with TCB is set to 16, and its channel number is 60. The cross-scale feature maps number $L = 4$ is set to 4. The range of Top-k is set to $[50\%, 80\%]$. The number of supporting frames $2N = 4$. During training, we augment the data with random horizontal flips and rotation (90^{o}, 180^{o}, 270^{o}). We randomly crop LR video clips of size 64×64 with batch size 4 as input. Our model is trained by Adam op-

Video clips #Params	Ave (A, t)	TSP-FM 4.5+14.1M	TSP-DM 4.6+14.1M	TSP-FL 4.5+20.7M	TSP-ML 7.4+20.7M	TSP-DC 4.6+10.6M	TSP-MC 7.4+10.6M	ODPNet* 7.3M
Clip001	(0.7, 0.45) (0.7, 0.60) (0.8, 0.35) (0.8, 0.55) (0.9, 0.35) (0.9, 0.65) (1.0, 0.45) (1.0, 0.55)	21.64/0.8564	23.95/0.8996	21.64/0.8536	25.44/0.8912	24.00/0.9010	25.55/0.8950	28.96/0.9451
Clip002		24.91/0.9213	27.01/0.9335	24.89/0.9212	28.31/0.9356	27.11/0.9380	28.47/0.9399	31.92/0.9618
Clip003		20.88/0.8966	23.67/0.9257	20.87/0.8916	22.81/0.9165	23.69/0.9274	22.83/0.9225	24.22/0.9360
Clip004		29.30/0.8950	26.95/0.8828	29.23/0.8931	27.22/0.8787	26.91/0.8794	27.32/0.8802	29.54/0.9168
Clip005		26.46/0.9161	24.60/0.9074	26.47/0.9165	23.35/0.8889	24.61/0.9080	23.37/0.8904	37.28/0.9750
Clip006		21.97/0.8830	25.36/0.9128	21.97/0.8821	24.19/0.8940	25.41/0.9152	24.24/0.8962	31.93/0.9615
Clip007		18.59/0.8946	26.48/0.9565	18.60/0.8946	26.33/0.9559	26.52/0.9572	26.37/0.9569	33.13/0.9796
Clip008		24.30/0.8929	23.66/0.8901	24.30/0.8920	22.00/0.8784	23.72/0.8945	22.04/0.8812	30.99/0.9476
Clip009		18.41/0.8342	22.84/0.9052	18.40/0.8279	24.89/0.8987	22.97/0.9100	24.90/0.9057	31.61/0.9521
Clip010		20.47/0.7995	25.88/0.8799	20.45/0.8001	22.29/0.8569	26.12/0.8917	22.47/0.8727	31.65/0.9206

Table 2. Quantitative comparison (PSNR/SSIM) for $\times 4$ satellite VSR with TSP-based methods on ten video clips. Red and blue text indicates the best and second best performance respectively. Ave denotes the average result over all (A, t). * indicates our method.

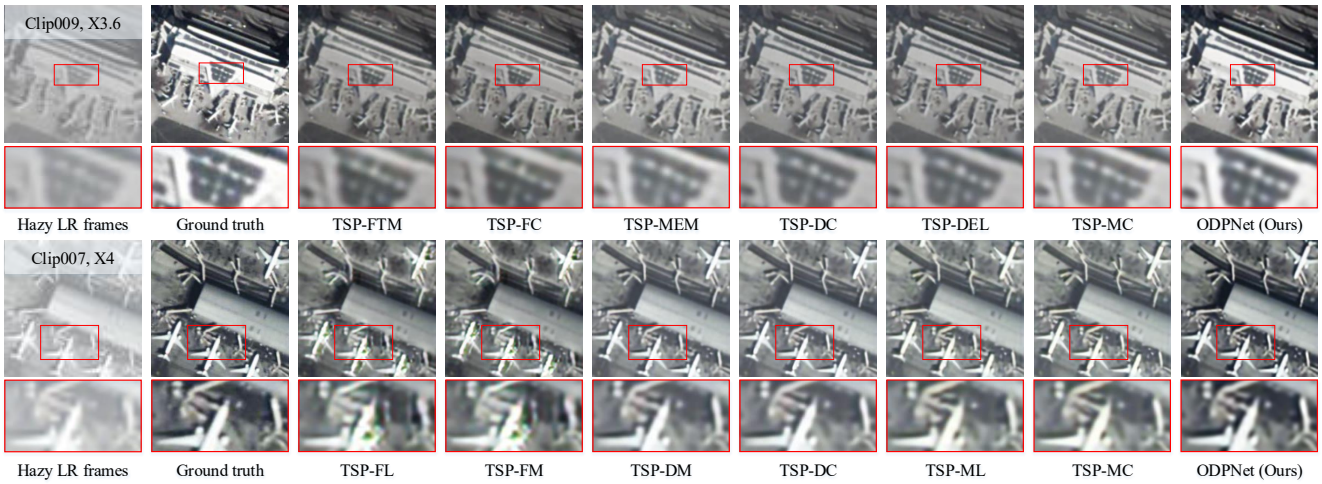

Figure 7. Qualitative comparison of ODPNet with other methods on scales $\times 3.6$ and $\times 4$. More details are marked with red boxes.

timizer [14] for 200K iterations. The initial learning rate is set to 2×10^{-4} and halved for every 30k iterations. LR video clips are obtained by double-degraded processing (by adding haze and reducing spatial resolution). In the process of haze synthesis, the airlight $A \in [0.7, 1]$ and the transmission $t \in [0.35, 0.6]$ are randomly sampled. In the process of reducing resolution, we use the bicubic kernel.

4.2. Comparison with satellite VSR methods

Due to the lack of a unified framework for hazy satellite VSR. Therefore, we use TSP methods as comparison methods. The dehazing methods include FFANet [25], DehazeFormer [30], MB-TaylorFormer [26], and the VSR methods include TDAN [31], EDVR [35], MSDTGP [40], LGTD [41], CSVSR [23]. Since some VSR methods cannot achieve continuous-scale VSR, we embed the existing continuous-scale upsampling methods Meta-SR [12] and LIFF [4] to complete it.

For a fair comparison on continuous-scale, we combine nine different TSP methods, including FFANet+TDAN-Meta-SR (TSP-FTM), DehazeFormer+EDVR-LIFF (TSP-DEL), MB-TaylorFormer+EDVR-Meta-SR (TSP-MEM), FFANet+EDVR-LIFF (TSP-FEL), DehazeFormer+TDAN-Meta-SR (TSP-DTM), MB-TaylorFormer+TDAN-LIFF (TSP-MTL), FFANet+CSVSR (TSP-FC), DehazeFormer+CSVSR (TSP-DC), MB-TaylorFormer+CSVSR (TSP-MC). As shown in Table 1, ODPNet outperforms all TSP methods and achieves a good balance between model parameters and reconstruction quality on all scale factors. In addition, we combine six different TSP methods for $\times 4$ comparison, including FFANet+MSDTGP (TSP-FM), DehazeFormer+MSDTGP (TSP-DM), FFANet+LGTD (TSP-FL), MB-TaylorFormer+LGTD (TSP-ML), DehazeFormer+CSVSR (TSP-DC), MB-TaylorFormer+CSVSR (TSP-MC). Comparison results are shown in Table 2, we can find that our ODPNet achieves the best performance

Models	Ave (A, t)	$\times$2.1	$\times$3.2	$\times$4
w/o PFSD	(0.7, 0.35) (0.8, 0.45) (0.9, 0.55) (1.0, 0.60)	32.13/0.9892	29.93/0.9740	28.34/0.9424
ODPNet		**33.65/0.9916**	**31.84/0.9767**	**28.63/0.9448**

Table 3. Ablation studies of the proposed PFSD on Clip001.

Model #	Ave (Clips)	Temporal-SA	Spatial-SA	$\times$4
1	Clip001-010	$\times$	$\times$	28.53/0.9418
2		✓	$\times$	28.86/0.9425
ODPNet		✓	✓	**29.30/0.9464**

Table 4. Ablation studies of the proposed SSA on haze (0.7, 0.35).

Figure 8. Visual results of ODPNet on real hazy satellite video. Top row: Hazy satellite video frame. Bottom row: Corresponding restoration results ($\times$4). More details are marked with red boxes.

with fewer model parameters.

We also perform a qualitative comparison with other TSP methods. As shown in Fig. 7, our method produces higher quality HR video, including correct color, finer details, and sharper edges. The dehazing and SR results of the comparison methods still have some defects, such as light haze coverage, color deviation, texture disorder, and unclear edges. In addition, we evaluated our ODPNet on real-world LR hazy satellite videos captured by the International Space Station (ISS). As shown in Fig. 8, our method ODPNet successfully restores and reconstructs hazy LR satellite videos, which further validates the effectiveness and robustness of our method in practical real-world scenarios.

4.3. Ablation Studies

First, to demonstrate the effectiveness of our prior-based feature soft dehazing (PFSD), we designed a baseline model without (w/o) predicting K and B (w/o PFSD) to achieve feature-level dehazing. As shown in Table 3, we can observe that our PFSD can effectively improve network performance and gain **+1.91**dB over w/o PFSD on $\times$3.2.

Second, to verify the effectiveness of our proposed SSA, we designed two models. Model 1 abandons all self-attention (SA), and Model 2 only performs temporal SA. As shown in Table 4, Model 2 can use temporal SA to effectively utilize complementary information in the temporal dimension, but cannot use spatial SA to propagate complementary information to the spatial dimension, and only surpasses Model 1 by **+0.33**dB. ODPNet can use SSA to fully utilize complementary information in the temporal and spatial dimensions and gain **+0.77**dB over Model 1.

Models	Ave (Scales)	(0.7, 0.35)	(0.8, 0.45)	(0.9, 0.55)
w/o IA	$\times$(2.1,2.7,3.2,4)	29.93/0.9697	29.76/0.9716	29.69/0.9718
ODPNet		**30.67/0.9741**	**31.44/0.9746**	**31.78/0.9748**

Table 5. Ablation studies of the proposed IA on Clip001.

Top-k	Ave (Clips)	$\times$2.1	$\times$3.2	$\times$4
k=50%	Clip001-010	36.13/0.9898	33.28/0.9717	31.58/0.9441
k=[50%,66%]		36.85/0.9897	34.27/0.9739	31.54/0.9486
k=[50%,75%]		35.97/0.9897	34.04/0.9743	31.98/0.9492
k=[50%,80%] (Ours)		37.02/0.9906	34.83/0.9760	32.27/0.9510
w/o Top-k		36.05/0.9881	33.75/0.9732	31.58/0.9484

Table 6. Ablation studies of the proposed Top-k of TST on haze (1.0, 0.6).

Third, to prove the effectiveness of our introduction of IA in TCB, we perform a comparison experiment between ODPNet and ODPNet w/o IA. As shown in Table 5, by aggregating the features from the spatial SA branch and channel SA branch, ODPNet can more effectively mine the efficient feature representation of detailed texture in satellite videos, thereby achieving higher-quality reconstruction.

Finally, to verify the effectiveness of the cross-scale Top-k selection mechanism, we further discussed the selection of the k value of Top-k, and the experimental results are shown in Table 6. w/o means that the Top-k selection is not performed, and all cross-scale features are used to complete the mapping of features to pixel values. In addition, we found that if k is selected to be a small value (50%), some key features may be missed, resulting in reduced network performance. When the k is set to $[50\%, 80\%]$, ODPNet achieves the best performance.

5. Conclusion

In this paper, we propose a novel network (ODPNet) to learn the optimal joint degradation pattern for hazy satellite video restoration and continuous-scale VSR. First, we design a prior-based feature soft dehazing module, which removes haze interference at the feature level and achieves effective extraction of haze-free features. Second, we propose a spatiotemporal self-attention, which can comprehensively mine the complementary information to achieve an effective fusion of spatiotemporal features. Third, we develop a tri-branch cross-aggregation block to enhance feature representations by effectively aggregating contextual information. Finally, we propose a cross-scale feature Top-k selection Transformer, which aims to select the most relevant Top-k feature latent codes at multi-scale feature space to learn feature representations of satellite videos at arbitrary resolutions. Extensive experiments show the effectiveness and prominent performance of the proposed ODPNet.

References

[1] Guoling Bi, Guoliang Si, Yuchen Zhao, Biao Qi, and Hengyi Lv. Haze removal for a single remote sensing image using low-rank and sparse prior. *IEEE Trans. Geosci. Remote Sens.*, 60:1–13, 2022. 2

[2] Kelvin C.K. Chan, Shangchen Zhou, Xiangyu Xu, and Chen Change Loy. Basicvsr++: Improving video super-resolution with enhanced propagation and alignment. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 5962–5971, 2022. 1

[3] Xiang Chen, Yufeng Li, Longgang Dai, and Caihua Kong. Hybrid high-resolution learning for single remote sensing satellite image dehazing. *IEEE Geosci. Remote Sens. Lett.*, 19:1–5, 2022. 2

[4] Yinbo Chen, Sifei Liu, and Xiaolong Wang. Learning continuous image representation with local implicit image function. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 8628–8638, 2021. 2, 3, 7

[5] Zhaojie Chen, Qi Li, Huajun Feng, Zhihai Xu, and Yueting Chen. Nonuniformly dehaze network for visible remote sensing images. In *IEEE Conf. Comput. Vis. Pattern Recog. Worksh.*, pages 447–456, 2022. 2

[6] Kaichen Chi, Yuan Yuan, and Qi Wang. Trinity-net: Gradient-guided swin transformer-based remote sensing image dehazing and beyond. *IEEE Trans. Geosci. Remote Sens.*, 61:1–14, 2023.

[7] Jianhua Guo, Jingyu Yang, Huanjing Yue, Hai Tan, Chunping Hou, and Kun Li. Rsdehazenet: Dehazing network with channel refinement for multispectral remote sensing images. *IEEE Trans. Geosci. Remote Sens.*, 59(3):2535–2549, 2021. 2

[8] Yi Han, Ming Yin, Puhong Duan, and Pedram Ghamisi. Edge-preserving filtering-based dehazing for remote sensing images. *IEEE Geosci. Remote Sens. Lett.*, 19:1–5, 2022. 2

[9] Kaiming He, Jian Sun, and Xiaoou Tang. Single image haze removal using dark channel prior. *IEEE Trans. Pattern Anal. Mach. Intell.*, 33(12):2341–2353, 2010. 2

[10] Zongyao He and Zhi Jin. Dynamic implicit image function for efficient arbitrary-scale image representation. *arXiv preprint arXiv:2306.12321*, 2023. 3

[11] Zongyao He and Zhi Jin. Latent modulated function for computational optimal continuous image representation. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 26026–26035, 2024. 3

[12] Xuecai Hu, Haoyuan Mu, Xiangyu Zhang, Zilei Wang, Tieniu Tan, and Jian Sun. Meta-sr: A magnification-arbitrary network for super-resolution. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 1575–1584, 2019. 2, 7

[13] Xianyu Jin, Jiang He, Yi Xiao, and Qiangqiang Yuan. Learning a local-global alignment network for satellite video super-resolution. *IEEE Geosci. Remote Sens. Lett.*, 20:1–5, 2023. 2

[14] Diederik P Kingma and J Ba. Adam: A method for stochastic optimization. In *Int. Conf. Learn. Represent.*, 2015. 7

[15] Longxuan Kou, Shengyang Li, Jian Yang, and Yixuan Lv. Bsvos: Background interference suppression strategy for satellite video multi-object segmentation. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 2024. 1

[16] Jaewon Lee and Kyong Hwan Jin. Local texture estimator for implicit representation function. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 1929–1938, 2022. 2

[17] Jiayuan Li, Qingwu Hu, and Mingyao Ai. Haze and thin cloud removal via sphere model improved dark channel prior. *IEEE Geosci. Remote Sens. Lett.*, 16(3):472–476, 2019. 2

[18] Shan Liang, Tao Gao, Ting Chen, and Peng Cheng. A remote sensing image dehazing method based on heterogeneous priors. *IEEE Trans. Geosci. Remote Sens.*, 62:1–13, 2024. 2

[19] Huan Liu and Yanfeng Gu. Deep joint estimation network for satellite video super-resolution with multiple degradations. *IEEE Trans. Geosci. Remote Sens.*, 60:1–15, 2022. 2

[20] Huan Liu, Yanfeng Gu, Tengfei Wang, and Shengyang Li. Satellite video super-resolution based on adaptively spatiotemporal neighbors and nonlocal similarity regularization. *IEEE Trans. Geosci. Remote Sens.*, 58(12):8372–8383, 2020. 2

[21] Yunfan Lu, Zipeng Wang, Minjie Liu, Hongjian Wang, and Lin Wang. Learning spatial-temporal implicit neural representations for event-guided video super-resolution. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 1557–1567, 2023. 1

[22] S.G. Narasimhan and S.K. Nayar. Contrast restoration of weather degraded images. *IEEE Trans. Pattern Anal. Mach. Intell.*, 25(6):713–724, 2003. 3, 4

[23] Ning Ni and Libao Zhang. Deformable convolution alignment and dynamic scale-aware network for continuous-scale satellite video super-resolution. *IEEE Trans. Geosci. Remote Sens.*, 62:1–17, 2024. 2, 4, 7

[24] Zhaoliang Pi, Licheng Jiao, Fang Liu, Xu Liu, Lingling Li, Biao Hou, and Shuyuan Yang. Very low-resolution moving vehicle detection in satellite videos. *IEEE Trans. Geosci. Remote Sens.*, 60:1–17, 2022. 1

[25] Xu Qin, Zhilin Wang, Yuanchao Bai, Xiaodong Xie, and Huizhu Jia. Ffa-net: Feature fusion attention network for single image dehazing. In *AAAI*, pages 11908–11915, 2020. 1, 7

[26] Yuwei Qiu, Kaihao Zhang, Chenxi Wang, Wenhan Luo, Hongdong Li, and Zhi Jin. Mb-taylorformer: Multi-branch efficient transformer expanded by taylor formula for image dehazing. In *Int. Conf. Comput. Vis.*, pages 12802–12813, 2023. 7

[27] Jia Shao, Bo Du, Chen Wu, Mingming Gong, and Tongliang Liu. Hrsiam: High-resolution siamese network, towards space-borne satellite video tracking. *IEEE Trans. Image Process.*, 30:3056–3068, 2021. 1

[28] Huanfeng Shen, Chi Zhang, Huifang Li, Quan Yuan, and Liangpei Zhang. A spatial–spectral adaptive haze removal method for visible remote sensing images. *IEEE Trans. Geosci. Remote Sens.*, 58(9):6168–6180, 2020. 2

[29] Dilbag Singh, Manjit Kaur, Mohamed Yaseen Jabarulla, Vijay Kumar, and Heung-No Lee. Evolving fusion-based visibility restoration model for hazy remote sensing images using dynamic differential evolution. *IEEE Trans. Geosci. Remote Sens.*, 60:1–14, 2022. 2

[30] Yuda Song, Zhuqing He, Hui Qian, and Xin Du. Vision transformers for single image dehazing. *IEEE Trans. Image Process.*, 32:1927–1941, 2023. 1, 4, 7

[31] Yapeng Tian, Yulun Zhang, Yun Fu, and Chenliang Xu. Tdan: Temporally-deformable alignment network for video super-resolution. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 3360–3369, 2020. 1, 7

[32] Hang Wang, Xuanhong Chen, Bingbing Ni, Yutian Liu, and Jinfan Liu. Omni aggregation networks for lightweight image super-resolution. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 22378–22387, 2023. 2

[33] Pichao Wang, Xue Wang, Fan Wang, Ming Lin, Shuning Chang, Hao Li, and Rong Jin. Kvt: k-nn attention for boosting vision transformers. In *European conference on computer vision*, pages 285–302. Springer, 2022. 5

[34] Wenxiao Wang, Lu Yao, Long Chen, Binbin Lin, Deng Cai, Xiaofei He, and Wei Liu. Crossformer: A versatile vision transformer hinging on cross-scale attention. In *Int. Conf. Learn. Represent.*, 2022. 5

[35] Xintao Wang, Kelvin CK Chan, Ke Yu, Chao Dong, and Chen Change Loy. Edvr: Video restoration with enhanced deformable convolutional networks. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 0–0, 2019. 1, 7

[36] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Trans. Image Process.*, 13(4): 600–612, 2004. 6

[37] Min Wei and Xuesong Zhang. Super-resolution neural operator. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 18247–18256, 2023. 3

[38] Hanlin Wu, Ning Ni, and Libao Zhang. Learning dynamic scale awareness and global implicit functions for continuous-scale super-resolution of remote sensing images. *IEEE Trans. Geosci. Remote Sens.*, 61:1–15, 2023. 2, 3

[39] Hanlin Wu, Ning Ni, and Libao Zhang. Lightweight stepless super-resolution of remote sensing images via saliency-aware dynamic routing strategy. *IEEE Trans. Geosci. Remote Sens.*, 61:1–17, 2023. 3

[40] Yi Xiao, Xin Su, Qiangqiang Yuan, Denghong Liu, Huanfeng Shen, and Liangpei Zhang. Satellite video super-resolution via multiscale deformable convolution alignment and temporal grouping projection. *IEEE Trans. Geosci. Remote Sens.*, 60:1–19, 2022. 1, 2, 7

[41] Yi Xiao, Qiangqiang Yuan, Kui Jiang, Xianyu Jin, Jiang He, Liangpei Zhang, and Chia-wen Lin. Local-global temporal difference learning for satellite video super-resolution. *IEEE Trans. Circuit Syst. Video Technol.*, pages 1–1, 2023. 1, 2, 7

[42] Yi Xiao, Qiangqiang Yuan, Qiang Zhang, and Liangpei Zhang. Deep blind super-resolution for satellite video. *IEEE Trans. Geosci. Remote Sens.*, 2023. 2

[43] Cong Xu, Hua Qi, Yahui Zheng, and Shiyong Peng. Real-time moving vehicle detection in satellite video based on historical differential information and grouping features. *IEEE Transactions on Geoscience and Remote Sensing*, 2024. 1

[44] Xiaozhe Zhang, Fengying Xie, Haidong Ding, Shaocheng Yan, and Zhenwei Shi. Proxy and cross-stripes integration transformer for remote sensing image dehazing. *IEEE Trans. Geosci. Remote Sens.*, 62:1–15, 2024. 2

[45] Hengyuan Zhao, Xiangtao Kong, Jingwen He, Yu Qiao, and Chao Dong. Efficient image super-resolution using pixel attention. In *Eur. Conf. Comput. Vis.*, pages 56–72. Springer, 2020. 5

[46] Yitong Zheng, Jia Su, Shun Zhang, Mingliang Tao, and Ling Wang. Dehaze-aggan: Unpaired remote sensing image dehazing using enhanced attention-guide generative adversarial networks. *IEEE Trans. Geosci. Remote Sens.*, 60:1–13, 2022. 2

[47] Yanfei Zhong, Meng Shu, Zhenqi Liu, and Xiaoyan Lu. Spatio-temporal dual-branch network with predictive feature learning for satellite video object segmentation. *IEEE Trans. Geosci. Remote Sens.*, 60:1–15, 2022. 1

[48] Shangchen Zhou, Peiqing Yang, Jianyi Wang, Yihang Luo, and Chen Change Loy. Upscale-a-video: Temporal-consistent diffusion model for real-world video super-resolution. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 2535–2545, 2024. 1

[49] Qiqi Zhu, Xin Huang, and Qingfeng Guan. Tabctnet: Target-aware bilateral cnn-transformer network for single object tracking in satellite videos. *International Journal of Applied Earth Observation and Geoinformation*, 128:103723, 2024. 1