IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 30 May 2024, accepted 20 July 2024, date of publication 23 July 2024, date of current version 7 August 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3432709

RESEARCH ARTICLE

Video Desnower: An Adaptive Feature Fusion Understanding Video Desnowing Model With Deformable Convolution and KNN Point Cloud Transformer

YUXUAN LI AND LIN DAI
Beijing Institute of Technology, Beijing 100081, China

Corresponding author: Lin Dai (dailiu@bit.edu.cn)

This work was supported by the National Natural Science Foundation of China under Grant 62172039.

ABSTRACT The desnowing (snow removal) model is extensively utilized in various fields, including visual enhancement, security monitoring, and autonomous driving technology. Some previous work developed highly efficient models that primarily addressed single-image desnowing tasks. Simultaneously, the process of video desnowing holds significance in practical applications. There is only a limited amount of literature available on the topic of video desnowing, mainly utilizing predetermined knowledge rather than exploring deep learning technologies. Given the identified deficiency in current research, our study aims to improve upon existing video desnowing methodologies by introducing an innovative approach and filling the void of specialized datasets. Our contribution includes the development of a dataset tailored for the training and assessment of video desnowing models, as well as the creation of the Video-Denower model, which integrates adaptive feature fusion mechanisms. Video-Desnower employs sophisticated adaptive feature fusion methodologies to enhance desnowing efficacy through the comprehensive analysis of features across various scales. In contrast to single-image models, this particular model has the ability to analyze multiple frames within a video. Experiments on a video desnowing dataset show its exceptional capabilities. The code and dataset used in this study are available upon request. Interested researchers can contact us at liyux2001@163.com for access. Please include a brief description of your research interest and how you intend to use the data.

INDEX TERMS Computer vision, deep learning, video desnowing, feature fusion understanding.

I. INTRODUCTION

Recent advancements in computer vision have been notable due to the successful integration of deep learning technology. Deep learning techniques have been applied effectively in computer vision tasks, such as image classification [1], [2], [3], [4], [5], [6], [7], object detection [8], [9], [10], [11], [12], and semantic segmentation [13], [14], [15], [16], [17], [18], [19], [20]. Desnowing is a computer vision task that seeks to enhance visual clarity by eliminating snow noise from images or videos.

The process of removing snow can be mathematically characterized as Eq.(1) [21]. Let $\mathcal{I}(x)$ denote the snowy image at pixel location x, which can be modeled by the following equation:

$$\mathcal{I}(x) = \mathcal{K}(x)\mathcal{T}(x) + \mathcal{A}(x)(1 - \mathcal{T}(x)), \tag{1}$$

where $\mathcal{T}(x)$ represents the transmission map, indicating the proportion of the scene visible through the snow. $\mathcal{A}(x)$ signifies the atmospheric light, which affects the color and intensity due to the snow. $\mathcal{K}(x)$, representing the scene free of veiling effects from snow, can be decomposed as follows:

$$\mathcal{K}(x) = \mathcal{J}(x)(1 - \mathcal{Z}(x)\mathcal{R}(x)) + \mathcal{C}(x)\mathcal{Z}(x)\mathcal{R}(x), \tag{2}$$

The associate editor coordinating the review of this manuscript and approving it for publication was Gustavo Olague.

where $\mathcal{J}(x)$ is the clean image, devoid of any snow. $\mathcal{R}(x)$ is the binary mask indicating the locations of snow within the image. $\mathcal{Z}(x)$ denotes the chromatic aberration image, which captures the color distortions caused by snow. $\mathcal{C}(x)$ represents the color of the snow, accounting for the variations in snow's hue across the image.

Researchers proposed effective single-image desnowing solutions to remove snow from images. However, the utilization of video desnowing holds equal importance to single-image desnowing, particularly in light of the imperative need for ongoing processing of numerous video frames in practical environments. In light of the dearth of datasets available for the training and evaluation of video desnowing models in academic circles, we curated a dataset tailored to this objective in order to support the progression of research in video desnowing.

At present, current desnowing initiatives exhibit a deficiency in comprehensively grasping the amalgamation of snowflake characteristics. As a solution, we devised an innovative transformer that incorporates k-nearest-neighbors (KNN) point cloud processing capabilities to efficiently integrate and evaluate the multi-scale characteristics of snowflakes. Moreover, the integration of deformable convolutions with traditional convolutions was implemented to effectively tackle the irregular shapes of snowflake noise, ultimately improving the feature fusion and comprehension abilities of the model. To enhance the understanding of diverse characteristics, three perceptrons were integrated to autonomously modify the three K values of KNN, the comprehension weights for distinct K values, and the feature weights of the two convolution types. The integrated video desnowing model (Video-Desnower) was created and tested, showing excellent snow removal performance in videos.

Our work's contributions are outlined as follows:

- A dataset was generated for the purpose of training and assessing video desnowing models.
- A strong desnowing algorithm called Video-Desnower model was created with adaptive feature fusion.
- The exceptional desnowing performance of Video-Desnower was validated through experimental methods.

The following sections of this paper are organized as follows: Section II reviews pertinent literature, Section III provides a detailed analysis of Video-Desnower, Section IV presents the experiments and results, Section V discusses the ablation study, and Section VI offers a conclusion to the paper.

II. RELATED WORK

A. DESNOWING MODEL BASED ON PREDEFINED PRIORS

Prior research on desnowing primarily utilized models that were based on existing knowledge, frequently predetermined. Jérémie Bossu et al. posited that precipitation in the form of rain and snow could be modeled using a Gaussian distribution. They utilized the conventional Gaussian Mixture Model to distinguish between rain and snow, quantified the intensity of precipitation, and preserved images that were free from any obscuring rain or snow [22]. Wang et al. integrated image decomposition and dictionary learning methodologies to partition images into distinct layers, subsequently processing each layer independently to achieve the desnowing of images [23]. Huang et al. proposed a new methodology that incorporates a sparse image approximation module and an adaptive tolerance optimization module. Through iterative implementation, this approach successfully mitigated snowflake noise and produced images devoid of snow [24]. These methodologies rely on predetermined prior knowledge and lack the ability to autonomously adjust to alterations in snowflake attributes. When significant changes occur in properties such as snowflake characteristics, the model faces difficulties in efficiently carrying out desnowing procedures.

B. DEEP LEARNING IN COMPUTER VISION

In previous work, deep learning has been widely applied to computer vision tasks, such as medical image segmentation and classification, hyperspectral image classification, and object detection in remote sensing images. For instance, multi-task networks and class incremental learning methods have enhanced medical and hyperspectral image classification respectively [1], [2]. Interpretable models have improved neurological phenotyping [3], while semi-supervised learning and few-shot techniques have advanced image classification with limited labeled data [4], [5], [6]. Deep neural networks have automated endoscopic image classification [7], and weakly supervised object detection has benefited from methods selecting high-quality proposals [8]. In the realm of few-shot object detection, new networks tailored for remote sensing images have emerged [9]. Advancements also include automatic learning of object co-occurrence knowledge for remote sensing image detection [10], and the development of a multitask benchmark dataset for satellite video covering detection, tracking, and segmentation [11]. Methods for metal and living object detection in wireless charging systems have also been explored [12]. In 3D semantic segmentation, scene-adaptive approaches with multi-level enhancement have been proposed [13]. Domain adaptation techniques for semantic segmentation in remote sensing images have been improved with semantic-preserved generative adversarial networks [14].For RGB-thermal image segmentation, semantic-guided fusion networks have shown promise [15]. Semantic information extraction from various data forms in 3D point clouds has been investigated [16]. Semi-supervised semantic segmentation using cross-image semantic consistency has been developed [17], while RGB-T semantic segmentation has been enhanced through location, activation, and sharpening techniques [18]. Collaborative learning strategies for semi-supervised semantic segmentation have been introduced [19], and networks sharing modal memory and complementing morphological information for RGB-T urban scene segmentation have been proposed [20]. The removal of snow from videos is a significant computer vision task, and the implementation of deep learning methodologies is expected to enhance its efficacy.

C. DESNOWING MODEL BASED ON DEEP LEARNING TECHNOLOGY

Recently, desnowing models have incorporated deep learning technologies, allowing them to autonomously acquire knowledge of snowflake characteristics, thereby improving their adaptability and generalization abilities. Chen et al. developed an invertible neural network model with two asymmetric interactive pathways incorporating attention mechanisms, demonstrating the superior performance of the proposed model in single-image desnowing tasks [21]. Jaw et al. proposed a modular desnowing network architecture using generative adversarial network technology [25]. Cheng et al. proposed an adaptive residual network to capture the characteristics of snowflakes and locate the positions of snow, using a reconstruction network to generate snow-free images [26]. Chen et al. developed new snow models and proposed a desnowing algorithm that jointly recognizes the size and transparency of snow [27]. The success of these endeavors has motivated us to employ deep learning models in the pursuit of video desnowing objectives.

D. VIDEO DESNOWING MODEL

The works previously mentioned concentrate on desnowing individual images, yet video desnowing presents also a wide range of potential applications. There exist academic investigations pertaining to the process of video desnowing. Kim et al. utilized temporal correlations and low-rank matrix completion techniques to remove rain or snow streaks from video sequences [28]. Yang et al. proposed a novel video desnowing method based on adaptive snowflake detection and a patch-based Gaussian Mixture Model [29]. Tian et al. extracted a combination of static background and moving foreground along with falling snowflakes using global low-rank matrix decomposition, removed the falling snowflakes in front of moving objects using local low-rank decomposition, and generated snow-free videos by overlaying the moving foreground onto the static background [30]. It is imperative to recognize the scarcity of research on video desnowing, often utilizing antiquated methods and lacking sufficient incorporation of advanced deep learning techniques. Hence, it is imperative to develop novel deep learning models specifically designed for the task of video desnowing.

III. PROPOSED METHOD

Figure 1 depicts the architectural design of Video-Desnower, a model proposed for the purpose of video desnowing. Through the integration of MSBlock and K-Former, Video-Desnower demonstrates the capability to successfully merge and interpret snowflake attributes in a flexible and adaptive fashion. The MSBlock integrates feature maps obtained from deformable and conventional convolutions, establishes a perceptron for dynamically determining weights, and then applies these weights to combine the two feature maps.The K-Former model, a derivative of the Transformer architecture, employs a Feedforward design that integrates three KNN point cloud models. This methodology entails the utilization of data input into a perceptron for the purpose of ascertaining three adaptive K values, which are subsequently employed in another perceptron for the computation of three adaptive weights. The ultimate result of the Feedforward process is the weighted average of the three KNN point cloud models. This section will offer comprehensive explanations of both the MSBlock and K-Former modules.

A. MSBlock: AN ADAPTIVE FUSION BLOCK DESIGNED TO FACILITATE THE INTEGRATION OF 2 TYPES OF CONVOLUTIONAL FEATURE GRAPHS

1) MOTIVATION

In the task of image desnowing, the shape and distribution of snowflakes exhibit high randomness and irregularity, posing significant challenges to the model. Traditional convolutional neural networks (CNNs) perform well in handling smooth regions and regular structures, but their limitations become apparent when dealing with complex and irregular snowflake noise. To effectively address this issue, we designed a deformable convolution and conventional convolution integrated adaptive feature fusion module to enhance the model's ability to process snowflake noise.

Advantages of Deformable Convolution: deformable convolution is an enhanced convolution operation that introduces learnable offsets, allowing the convolutional kernel to adapt to irregular shapes and complex structures in the image. Compared to fixed-shape conventional convolution, deformable convolution can more flexibly capture the details and edges of snowflakes, thereby improving the model's ability to capture and remove snowflake noise.

Stability of Conventional Convolution: conventional convolution exhibits high stability and efficiency when handling smooth regions and regular structures. By combining conventional convolution with deformable convolution, we can enhance the model's capability to handle complex snowflake noise while ensuring processing efficiency. Conventional convolution provides a stable understanding of the overall image structure, while deformable convolution compensates for its shortcomings in dealing with irregular structures.

Adaptive Feature Fusion: to fully leverage the advantages of both deformable and conventional convolutions, we introduced an adaptive feature fusion mechanism. By learning weights, we can dynamically adjust the contributions of the two convolution operations based on the features of the input image. This adaptive fusion mechanism applies the most suitable convolution operation to different regions of the image, thereby enhancing the model's overall desnowing performance.

The design of MSBlock aims to effectively address the complexity and randomness of snowflake noise by

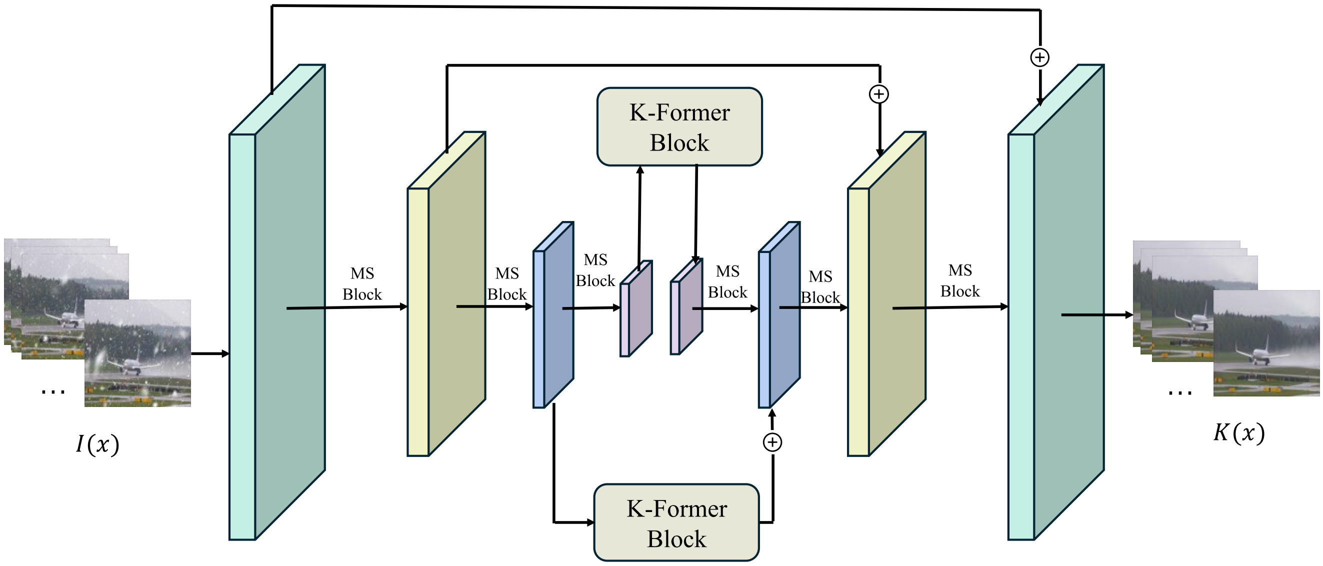

FIGURE 1. Structure of video-desnower. The core design of video-snower comprises the MSBlock and K-Former, with the primary advantage being the adaptive fusion-understanding of features.

combining the advantages of deformable and conventional convolutions and introducing an adaptive feature fusion mechanism. In this way, our model can more accurately capture and remove snowflake noise, achieving cleaner and higher-quality desnowing effects.

2) INTRODUCTION OF DEFORMABLE CONVOLUTION

The normal convolution operation in a neural network can be mathematically represented as:

$$y(p_0) = \sum_{p_n \in \mathcal{R}} w(p_n) \cdot x(p_0 + p_n), \tag{3}$$

where $y(p_0)$ is the output feature map at location p_0, $x(p_0+p_n)$ represents the input feature map at location p_0 shifted by p_n, $w(p_n)$ is the weight associated with the kernel at displacement p_n, $\mathcal{R}$ denotes the set of locations in the convolutional kernel.

Deformable convolution introduces an additional degree of freedom that allows the convolutional grid to adapt to the input feature map dynamically. It is defined as:

$$y(p_0) = \sum_{p_n \in \mathcal{R}} w(p_n) \cdot x(p_0 + p_n + \Delta p_n), \tag{4}$$

where Δp_n represents the learnable offset for the convolutional grid, other terms have the same meaning as in conventional convolution.

While normal convolution operates on a fixed grid pattern, deformable convolution adjusts the positions of the convolutional kernels during the learning process. This adaptability allows deformable convolution to handle geometric and spatial transformations more effectively than conventional convolution, making it especially useful in complex visual tasks where the alignment of features is critical.

By combining deformable convolution with normal convolution, we can more effectively handle snowflakes of varying shapes.

3) CONCRETE IMPLEMENTATION OF MSBlock

The MSBlock is implemented with the following components:

- Conventional convolution (`conv1`) and deformable convolution (`de_conv`) are used to extract features from the input.
- A fully connected network (`weight_net`) computes adaptive weights that control the blending of features from the two convolution types.

Given an input x from the dataset, the feature extraction and fusion process can be described as follows:

- Feature Extraction. Features are extracted using:

$$l_x = \texttt{conv1}(x), \tag{5}$$

$$\texttt{offset} = \texttt{conv2}(x), \tag{6}$$

$$g_x = \texttt{de\_conv}(x, \texttt{offset}), \tag{7}$$

where l_x and g_x represent features from conventional and deformable convolutions, respectively.

- Learning Weights. The perceptron within `weight_net` computes the weights:

$$\mathbf{w} = \texttt{Softmax}\Bigg(\tag{8}$$

$$\texttt{Linear}\Bigg(\tag{9}$$

$$\texttt{Flatten}\big(\tag{10}$$

$$\texttt{AdaptiveAvgPool2d}(x)\big)\Bigg)\Bigg) \tag{11}$$

where $\mathbf{w} = [w_l, w_g]$ are the weights for the features from the respective convolutions.

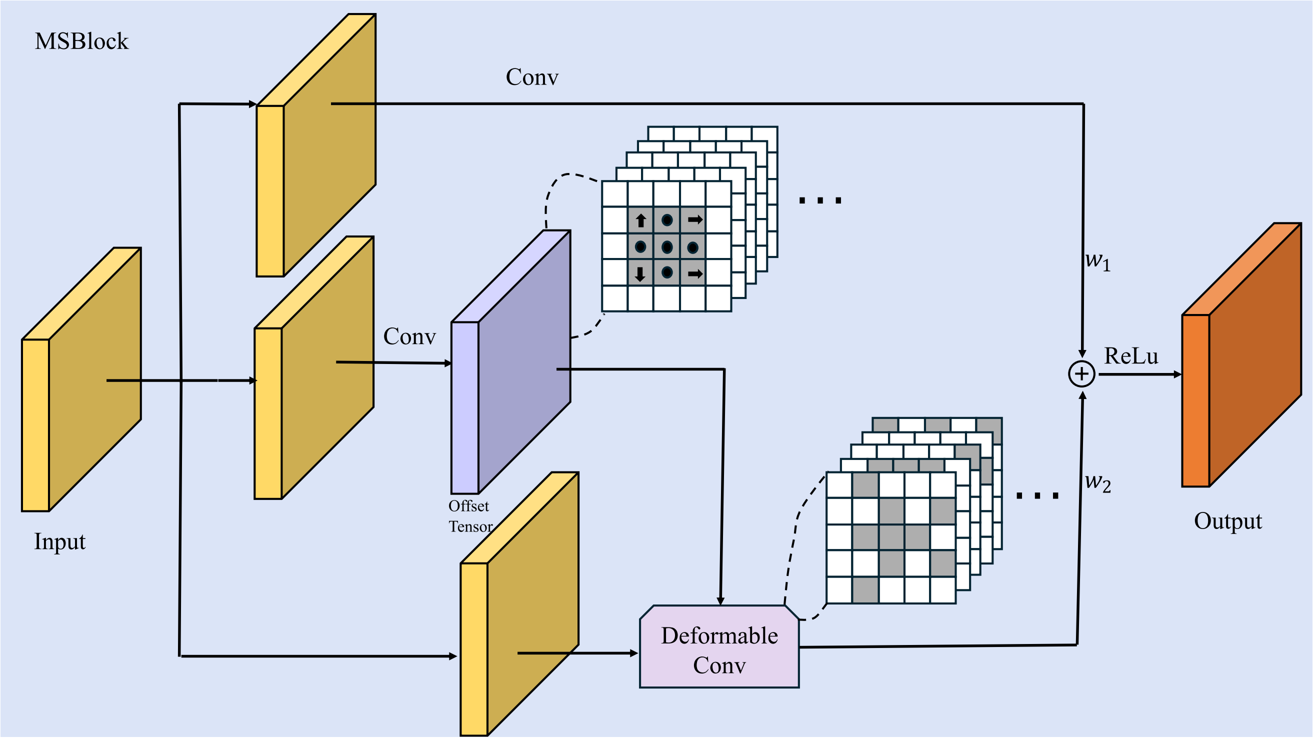

FIGURE 2. The design of MSBlock. The MSBlock fuses features from deformable and regular convolutions according to weights, where weights w_1 and w_2 are adaptively determined by a perceptron.

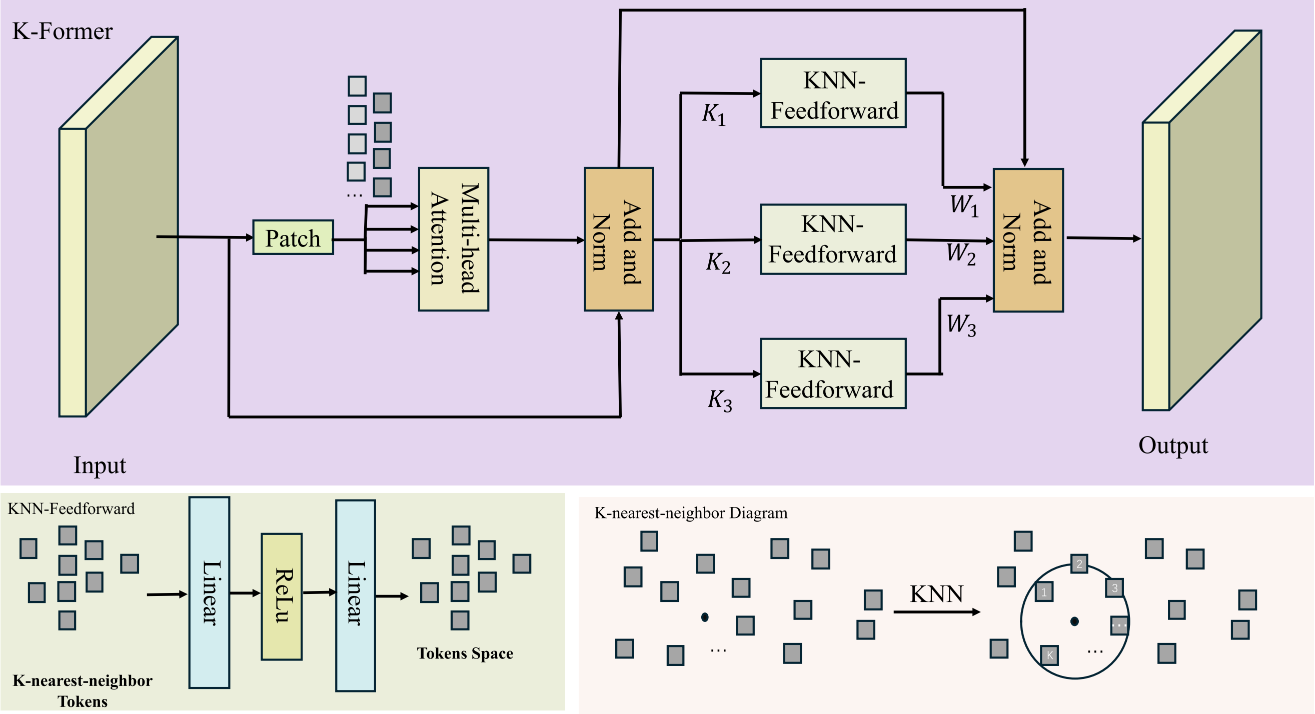

FIGURE 3. The design of K-former. The diagram contains K-Former's overall architecture and component KNN-Feedforward as well as an illustration of KNN technology. The K-Former utilizes the K-nearest neighbors (KNN) point cloud model to achieve an adaptive understanding and fusion of multi-scale features. The K values k_1 and k_2 are adaptively determined by one perceptron, while the weights w_1, w_2 and w_3 are adaptively obtained through another perceptron.

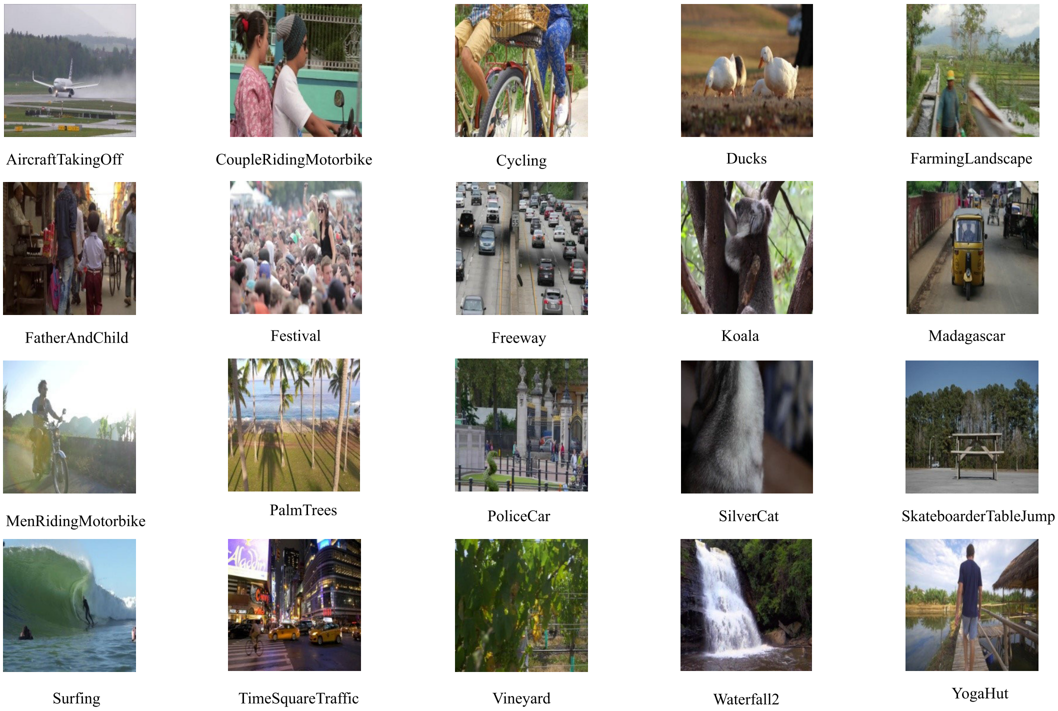

FIGURE 4. Presentation of the ground truth videos of the constructed dataset. One frame is selected for each video and each frame is labeled with a video name.

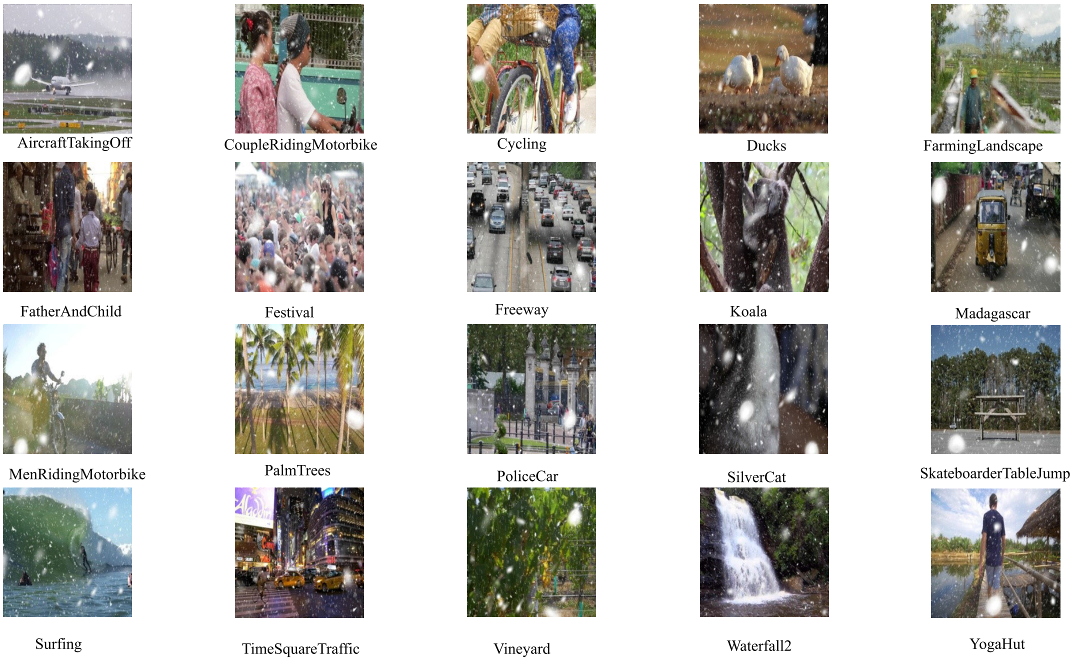

FIGURE 5. Video presentation of the of the constructed dataset with snowflake noise added. One frame is selected for each video and each frame is labeled with a video name.

- Feature Fusion. The adaptive fusion of the extracted features is performed as:

$$x_{\text{out}} = \texttt{ReLU}(w_l \cdot l_x + w_g \cdot g_x), \tag{12}$$

  where x_{out} is the output after feature blending.

The structure of the MSBlock is shown in Figure 2.

B. K-FORMER: A REFINED TRANSFORMER MODEL FOR ENHANCING COMPREHENSION OF ADAPTIVE KNN FUSION

1) MOTIVATION

K-Former is an advanced version of the Transformer model, designed to incorporate adaptive KNN point cloud models

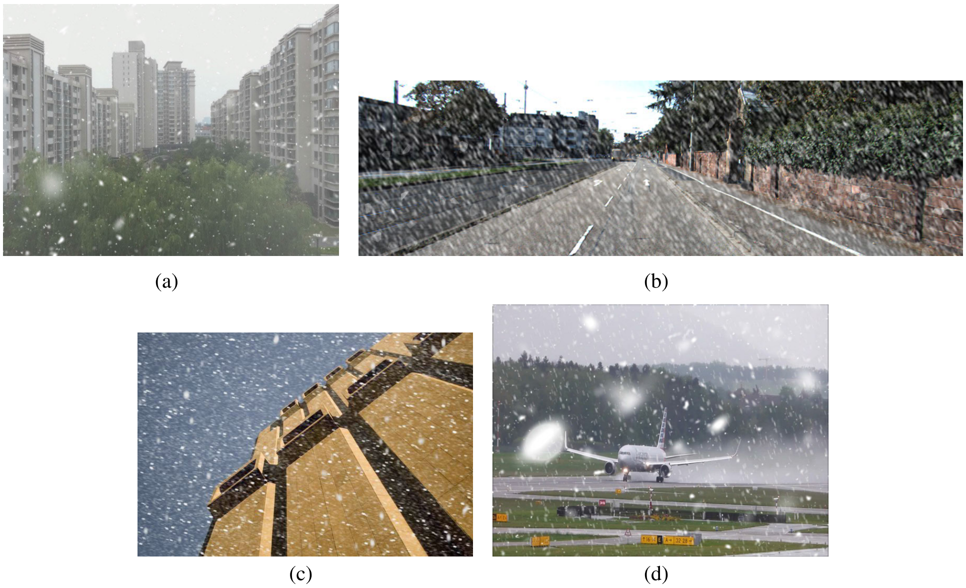

FIGURE 6. Comparison of snowflake noise levels across different datasets demonstrates that our custom dataset exhibits significantly larger and more complex snowflake noise, making the desnowing task more challenging. (a) SRRS [27], (b) SnowKITTI [31], (c) Snow100K [32], (d) Our video desnowing dataset.

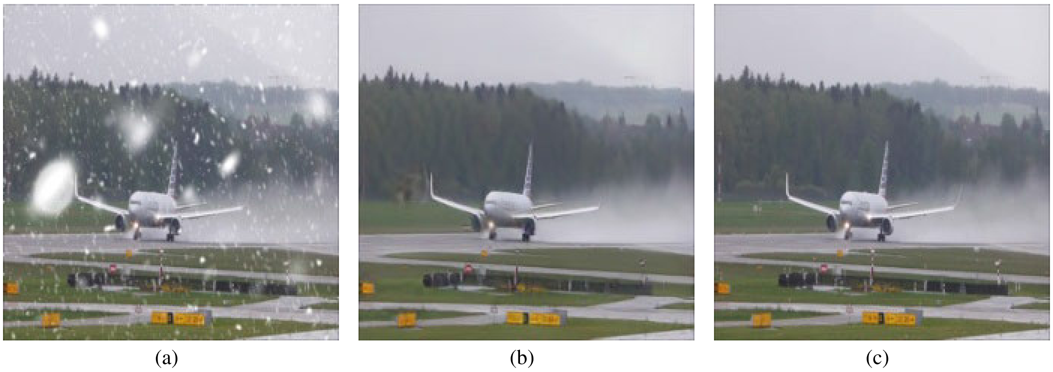

FIGURE 7. Display of desnowing effects for one frame in the dataset. (a) Model input (b) Model output (c) Ground truth.

within its architecture. This innovative structure allows the model to dynamically adapt to input features, enhancing its effectiveness in tasks requiring nuanced spatial understanding.

In the task of image desnowing, the shapes and distributions of snowflakes are highly random and irregular, posing significant challenges to the model's spatial understanding capabilities. Traditional convolutional neural networks, while effective at handling smooth regions and regular structures, exhibit limitations when dealing with complex and irregular snowflake noise. To effectively address this issue, we designed the K-Former model, which incorporates adaptive KNN point cloud models to enhance the model's ability to handle snowflake noise.

Advantages of Adaptive KNN: the adaptive KNN point cloud model dynamically selects the k-nearest points, allowing the model to adjust adaptively based on the input image features. This adaptive mechanism enables the model to better capture the details and edges of snowflakes, improving its ability to handle complex and irregular snowflake noise. Compared to a fixed number of nearest points, the adaptive KNN can apply the most suitable number of nearest points in different image regions, thereby enhancing the model's spatial understanding capabilities.

Enhancements to the Transformer: K-Former enhances the traditional Transformer model by incorporating the adaptive KNN point cloud model, endowing it with stronger spatial feature understanding capabilities. The Transformer model excels at capturing global features, and with the introduction of adaptive KNN, the model becomes more flexible in handling local complex structures, especially in cases of high randomness and irregularity, such as snowflake noise.

Adaptive Feature Fusion: to fully leverage the advantages of both the adaptive KNN point cloud model and the Transformer, we introduced an adaptive feature fusion mechanism. By learning weights, K-Former can dynamically adjust the

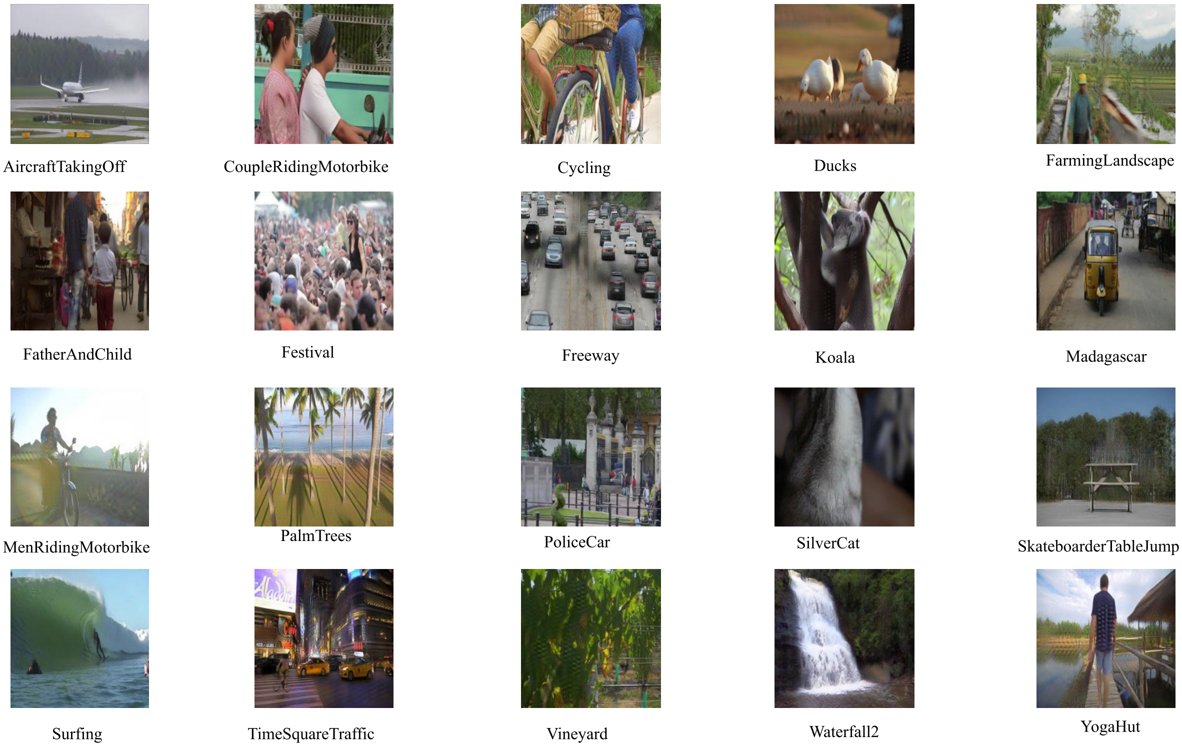

FIGURE 8. Comprehensive desnowing outcomes. One frame is selected for each video and each frame is labeled with a video name.

TABLE 1. Evaluation metrics for video-desnower desnowing performance.

Video Name	Number of Frames	Average SSIM	Average PSNR
AircraftTakingOff	300	0.99	33.75
CoupleRidingMotorbike	264	0.99	29.75
Cycling	249	0.98	26.13
Ducks	300	0.99	31.38
FarmingLandscape	253	0.99	29.74
FatherAndChild	250	0.98	29.07
Festival	250	0.99	30.52
Freeway	210	0.94	23.25
Koala	175	0.99	27.94
Madagascar	215	0.99	31.27
MenRidingMotorbike	300	0.99	30.97
PalmTrees	300	0.98	27.71
PoliceCar	239	0.99	30.56
SilverCat	227	0.99	32.52
SkateboarderTableJump	120	0.93	24.72
Surfing	143	0.99	33.89
TimeSquareTraffic	300	0.97	25.27
Vineyard	299	0.97	27.94
Waterfall	150	0.98	25.18
YogaHut	250	0.99	28.12
Overall Average	-	**0.98**	**29.13**

contributions of adaptive KNN and Transformer based on the input image features. This adaptive fusion mechanism allows the model to apply the most appropriate feature processing method in different image regions, thereby enhancing the overall desnowing performance.

The design of K-Former aims to effectively address the complexity and randomness of snowflake noise by combining the advantages of the adaptive KNN point cloud model and the Transformer, along with the introduction of an adaptive feature fusion mechanism. This approach enables our model to more accurately capture and remove snowflake noise, achieving cleaner and higher-quality desnowing results.

TABLE 2. Comparison of desnowing performance on our video dataset.

Model	Overall video frames	
	PSNR	SSIM
DesnowNet [32]	19.93	0.84
CycleGAN [35]	20.13	0.82
JSTASR [27]	27.10	0.85
DDMSNet [31]	27.79	0.87
HDCW-Net [36]	28.81	0.90
SMGARN [26]	29.10	0.93
Video-Desnower (Ours)	**29.13** ↑	**0.98** ↑

2) MODEL ARCHITECTURE

The K-Former modifies the traditional Transformer by integrating KNN point cloud models into the Feedforward network. It uses perceptrons to adaptively determine the best parameters for processing the input data. The overall structure of the K-Former is shown in Figure 3.

3) KNN-FEEDFORWARD

The Feedforward section of the K-Former is uniquely designed to include a KNN point cloud model, which is utilized to enhance the model's feature processing capabilities. The operations can be mathematically described as follows:

a: ADAPTIVE K-VALUE SELECTION

Three K values for the KNN model are adaptively determined by a perceptron:

$$K_i = int(\sigma(\mathbf{W}_k \cdot \mathbf{x} + \mathbf{b}_k)), \quad i = 1, 2, 3, \tag{13}$$

TABLE 3. Ablation study on video-desnower components (Part 1).

Component | Video name	AircraftTakingOff		CoupleRidingMotorbike	
	SSIM	PSNR	SSIM	PSNR
Full Model	0.99	33.75	0.99	29.75
Without Deformable Convolution	0.97	31.80	0.97	27.50
Without KNN	0.98	32.10	0.98	28.00
Without Adaptive Feature Fusion	0.98	32.50	0.98	28.25
Without Transformer Enhancements	0.97	31.50	0.96	27.00
Component	Cycling		Ducks	
	SSIM	PSNR	SSIM	PSNR
Full Model	0.98	26.13	0.99	31.38
Without Deformable Convolution	0.96	24.30	0.97	29.20
Without KNN	0.97	25.10	0.98	30.00
Without Adaptive Feature Fusion	0.97	25.40	0.98	30.30
Without Transformer Enhancements	0.95	24.00	0.97	28.50

Component | Video name	FarmingLandscape		FatherAndChild	
	SSIM	PSNR	SSIM	PSNR
Full Model	0.99	29.74	0.98	29.07
Without Deformable Convolution	0.97	27.20	0.96	27.10
Without KNN	0.98	28.00	0.97	28.20
Without Adaptive Feature Fusion	0.98	28.25	0.97	28.30
Without Transformer Enhancements	0.96	27.00	0.95	27.00

where σ denotes the sigmoid function, *int* is the integer function, $\mathbf{W}_k$ represents the weight matrix, $\mathbf{b}_k$ is the bias vector, and K_i are the adaptive K values.

b: WEIGHT ADAPTATION FOR KNN POINT CLOUDS

Another perceptron adjusts the weights for combining the outputs from the three KNN models:

$$w_i = \text{Softmax}(\mathbf{W}_w \cdot \mathbf{x} + \mathbf{b}_w), \quad i = 1, 2, 3, \tag{14}$$

where $\mathbf{W}_w$ and $\mathbf{b}_w$ are the weight matrix and bias vector for the weight adaptation perceptron, respectively, and w_i are the adaptive weights for each KNN model.

c: KNN POINT CLOUD MODEL

The KNN point cloud model operates by selecting the K-nearest neighbors of each data point in the feature space. Mathematically, this can be represented as:

$$\mathbf{K}NN_i(\mathbf{x}, K_i) = \text{TopK}\left(\|\mathbf{x} - \mathbf{x}_j\|, K_i\right), \tag{15}$$

where $\mathbf{x}_j$ are the data points in the training set, $\|\cdot\|$ denotes the Euclidean distance, and TopK retrieves the K smallest distances for the i-th KNN model.

d: OUTPUT COMPUTATION

The final output of the Feedforward network is the weighted average of the outputs from the three KNN point cloud models:

$$\mathbf{y} = \sum_{i=1}^{3} w_i \cdot \mathbf{K}NN_i(\mathbf{x}, K_i), \tag{16}$$

where $\mathbf{K}NN_i$ represents the output from the i-th KNN point cloud model using the adaptive K_i.

IV. EXPERIMENTS

A. DATASETS

Due to the scarcity of video desnowing datasets, we created a specialized dataset for this purpose. The dataset comprises 20 videos depicting snow-free scenes, each paired with a corresponding version that has been artificially augmented with snow. Each video sequence consists of a range of 100 to 400 consecutive frames. We provided this dataset in our research paper to help other researchers explore and improve video desnowing techniques. For each ground truth video within the dataset, a single frame was chosen and depicted in Figure 4.

For each corresponding video incorporating snowflake noise, a single frame was selected and presented in Figure 5.

To improve desnowing algorithms, our video model includes a lot of snowflake noise for added complexity and adaptability. Figure 6 compares snowflake density in a single frame from our dataset with other desnowing datasets, showing that our dataset has denser and larger snowflake noise. Video-Desnower effectively removes snowflakes from our dataset.

B. IMPLEMENTATION DETAILS

We optimized performance by using specific input and training configurations in our experiments. The input images were resized to 224 × 224 pixels, and each sample contained 5 frames. During the training phase, we used the Adam optimizer with an initial momentum $\beta_1 = 0$ and $\beta_2 = 0.99$. The initial learning rate was set to 0.0001. Training was conducted with a batch size of 8, and two worker threads were set up to load the data. Our model was implemented on the PyTorch [33] platform using a single RTX 3090 GPU, 16GB of RAM and an Intel i7 processor. We will provide all the code and the dataset we created.

C. EXPERIMENTAL RESULTS PRESENTATION

We utilized the Video-Desnower model to desnow 20 videos in our dataset, with all output videos and frames available. The findings illustrate the model's ability to enhance video frames that are significantly distorted by snowflake noise. Figure 7 depicts a representative video frame for visual reference.

The comprehensive desnowing outcomes are depicted in Figure 8.

TABLE 4. Ablation study on video-desnower components (Part 2).

Component | Video name	Festival		Freeway	
	SSIM	PSNR	SSIM	PSNR
Full Model	0.99	30.52	0.94	23.25
Without Deformable Convolution	0.97	28.60	0.92	22.10
Without KNN	0.98	29.00	0.93	22.80
Without Adaptive Feature Fusion	0.98	29.20	0.93	23.00
Without Transformer Enhancements	0.96	28.00	0.91	21.90
Component | Video name	Koala		Madagascar	
	SSIM	PSNR	SSIM	PSNR
Full Model	0.99	27.94	0.99	31.27
Without Deformable Convolution	0.97	26.50	0.97	29.40
Without KNN	0.98	27.20	0.98	30.00
Without Adaptive Feature Fusion	0.98	27.30	0.98	30.20
Without Transformer Enhancements	0.96	26.00	0.96	28.90

Component | Video name	MenRidingMotorbike		PalmTrees	
	SSIM	PSNR	SSIM	PSNR
Full Model	0.99	30.97	0.98	27.71
Without Deformable Convolution	0.97	29.00	0.96	26.20
Without KNN	0.98	29.50	0.97	26.80
Without Adaptive Feature Fusion	0.98	29.60	0.97	27.00
Without Transformer Enhancements	0.96	28.00	0.95	25.50

TABLE 5. Ablation study on video-desnower components (Part 3).

Component | Video name	PoliceCar		SilverCat	
	SSIM	PSNR	SSIM	PSNR
Full Model	0.99	30.56	0.99	32.52
Without Deformable Convolution	0.97	28.70	0.97	30.50
Without KNN	0.98	29.20	0.98	31.00
Without Adaptive Feature Fusion	0.98	29.30	0.98	31.20
Without Transformer Enhancements	0.96	28.00	0.96	29.80
Component | Video name	SkateboarderTableJump		Surfing	
	SSIM	PSNR	SSIM	PSNR
Full Model	0.93	24.72	0.99	33.89
Without Deformable Convolution	0.91	23.50	0.97	31.90
Without KNN	0.92	24.00	0.98	32.50
Without Adaptive Feature Fusion	0.92	24.10	0.98	32.70
Without Transformer Enhancements	0.90	22.50	0.96	30.80

Component	TimeSquareTraffic		Vineyard	
	SSIM	PSNR	SSIM	PSNR
Full Model	0.97	25.27	0.97	27.94
Without Deformable Convolution	0.95	23.90	0.95	26.30
Without KNN	0.96	24.50	0.96	27.00
Without Adaptive Feature Fusion	0.96	24.70	0.96	27.10
Without Transformer Enhancements	0.94	23.50	0.94	25.60

TABLE 6. Ablation study on video-desnower components (Part 4).

Component | Video name	Waterfall		YogaHut	
	SSIM	PSNR	SSIM	PSNR
Full Model	0.98	25.18	0.99	28.12
Without Deformable Convolution	0.96	24.00	0.97	26.50
Without KNN	0.97	24.50	0.98	27.00
Without Adaptive Feature Fusion	0.97	24.60	0.98	27.20
Without Transformer Enhancements	0.95	23.00	0.96	25.50

D. EVALUATION OF VIDEO-DESNOWER

Our model will be assessed through the utilization of SSIM and PSNR metrics. The importance of these metrics lies in their ability to provide quantitative measures of image quality and fidelity.

1) SSIM (STRUCTURAL SIMILARITY INDEX MEASURE) [34]

The Structural Similarity Index (SSIM) is used for measuring the similarity between two images. SSIM is based on the computation of three terms, namely the luminance term, the contrast term, and the structure term. The overall index is a multiplicative combination of the three terms:

$$\begin{aligned} \mathrm{SSIM}(x, y) = & \left[\frac{2\mu_x\mu_y + C_1}{\mu_x^2 + \mu_y^2 + C_1}\right] && (17) \\ & \cdot \left[\frac{2\sigma_{xy} + C_2}{\sigma_x^2 + \sigma_y^2 + C_2}\right] && (18) \\ & \cdot \left[\frac{\sigma_{xy} + C_3}{\sigma_x\sigma_y + C_3}\right] && (19) \end{aligned}$$

where μ_x, μ_y are the average values of x and y, σ_x, σ_y are the variance of x and y, and σ_{xy} is the covariance of x and y. C_1, C_2, and C_3 are constants to avoid division by zero.

2) PSNR (PEAK SIGNAL-TO-NOISE RATIO) [34]

The Peak Signal-to-Noise Ratio (PSNR) is most commonly used to measure the quality of reconstruction of lossy compression codecs (e.g., for image compression). The signal in this case is the original image, and the noise is the error introduced by compression. PSNR is usually expressed in terms of the logarithmic decibel scale:

$$\mathrm{PSNR} = 20 \cdot \log_{10}\left(\frac{\mathrm{MAX}_I}{\sqrt{\mathrm{MSE}}}\right), \tag{20}$$

TABLE 7. Overall average performance of video-desnower components.

Component	Average SSIM	Overall PSNR
Full Model	0.98	29.13
Without Deformable Convolution	0.96	27.50
Without KNN	0.97	28.00
Without Adaptive Feature Fusion	0.97	28.25
Without Transformer Enhancements	0.95	26.90

where MAX_I is the maximum possible pixel value of the image, and MSE is the Mean Squared Error between the original and the compressed image:

$$\mathrm{MSE} = \frac{1}{mn}\sum_{i=1}^{m}\sum_{j=1}^{n}(I(i,j) - K(i,j))^2, \qquad (21)$$

where I is the original image, K is the compressed image, and m and n are the dimensions of the images.

The SSIM and PSNR were were calculated in order to assess the level of resemblance between the model's output and the ground truth. Table 1 displays evaluation results. Video-Desnower shows strong desnowing performance despite high snowflake noise in the dataset.

E. COMPARED METHODS

The dataset utilized in this study is unique, as previous research on video desnowing has predominantly utilized datasets with lower levels of snowflake noise. In order to showcase the effectiveness of our model, we employed various open-source single-image desnowing models to process each frame within our video dataset. The resulting average SSIM and PSNR values for all frames are presented in the Table 2, providing evidence of the superior performance of our model. Of these, SSIM exhibits a notable enhancement.

V. ABLATION STUDY

To understand the contribution of each component in our Video-Desnower model, we conducted an ablation study by systematically removing or altering specific components and evaluating the impact on desnowing performance.

From Tables 3 to 7, we can observe the following detailed impacts of each component on the performance for different videos:

- **Deformable Convolution**: Removing deformable convolution results in a significant drop in both SSIM and PSNR across all videos. For example, in the AircraftTakingOff video, the SSIM decreased from 0.99 to 0.97 and PSNR from 33.75 to 31.80. This highlights the importance of deformable convolution in handling the irregular shapes of snowflakes. Similarly, in the Ducks video, SSIM dropped from 0.99 to 0.97 and PSNR from 31.38 to 29.20. The deformable convolution effectively captures the irregular and dynamic nature of snowflakes, providing the model with enhanced flexibility to adapt to varying snowflake shapes and densities.
- **KNN**: The absence of KNN shows a decrease in performance, indicating that KNN effectively enhances the model's ability to capture spatial relationships. For instance, in the CoupleRidingMotorbike video, the SSIM decreased from 0.99 to 0.98 and PSNR from 29.75 to 28.00. In the Koala video, the SSIM decreased from 0.99 to 0.98 and PSNR from 27.94 to 27.20. The adaptive KNN component is crucial for maintaining spatial coherence and accurately identifying snowflakes by considering the local neighborhood relationships within the image.
- **Adaptive Feature Fusion**: Without adaptive feature fusion, the model's performance declines, demonstrating the benefit of dynamically adjusting feature contributions. For the Cycling video, the SSIM dropped from 0.98 to 0.97 and PSNR from 26.13 to 25.40. In the SilverCat video, the SSIM decreased from 0.99 to 0.98 and PSNR from 32.52 to 31.20. Adaptive feature fusion allows the model to selectively integrate features from different layers and scales, enhancing the model's ability to handle varying snowflake sizes and intensities across different image regions.
- **Transformer Enhancements**: Removing transformer enhancements leads to a noticeable decrease in performance, suggesting their role in capturing global features. For the Freeway video, the SSIM decreased from 0.94 to 0.91 and PSNR from 23.25 to 21.90. Similarly, in the TimeSquareTraffic video, the SSIM dropped from 0.97 to 0.94 and PSNR from 25.27 to 23.50. The transformer enhancements enable the model to capture long-range dependencies and global contextual information, which are essential for accurately reconstructing complex scenes with heavy snow noise.

The overall results of Ablation study can be summarized in Figure 9 and 10. The ablation study confirms that each component is integral to the superior desnowing performance of the Video-Desnower model. The detailed comparison across different videos shows consistent positive impacts from all the components, underlining the effectiveness of each module in enhancing the model's ability to remove snow noise and produce high-quality, clear frames.

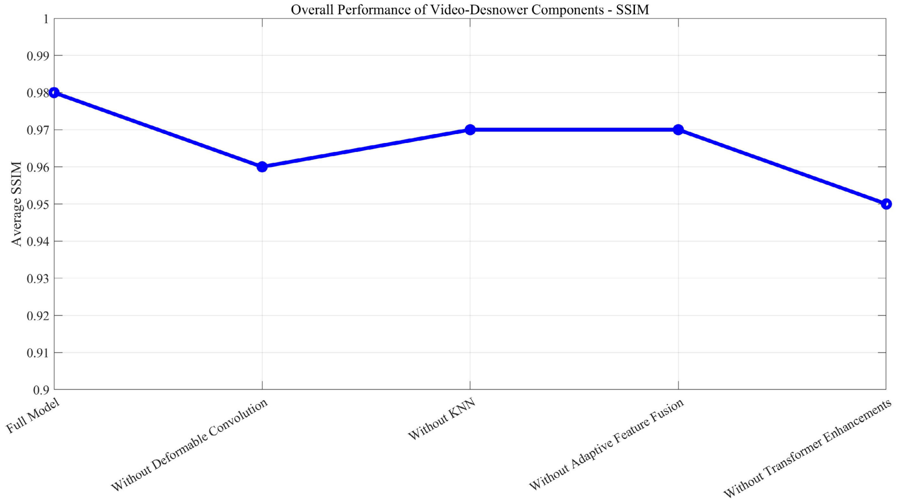

FIGURE 9. Overall results of ablation study (SSIM).

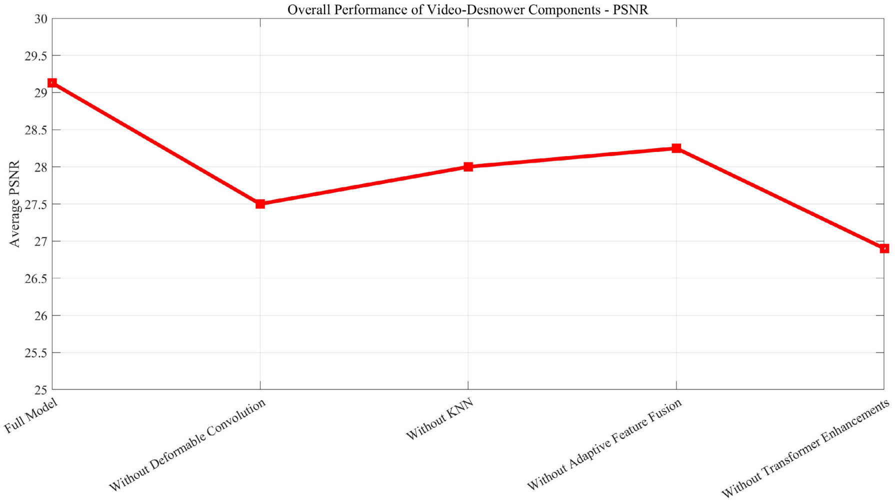

FIGURE 10. Overall results of ablation study (PSNR).

VI. CONCLUSION

Due to the scarcity of existing research on video desnowing solutions, we put forth a proposal for the utilization of a sophisticated deep learning model known as Video-Desnower. This model effectively eliminated snow interference from several consecutive video frames through the utilization of adaptive feature fusion. We tested our video desnowing model by adding snow noise to 20 real videos, creating a new dataset. Our analyses demonstrated that Video-Desnower consistently displayed impressive desnowing efficacy, even when faced with significant snowflake distortion in the dataset videos. The results of our study have significant practical implications for improving visual clarity, enhancing safety surveillance, facilitating autonomous driving, and a range of other potential applications. The dataset we offer is suitable for additional research on video desnowing. Future studies may use adaptive feature fusion in a wider variety of deep learning applications.

REFERENCES

[1] Y. Ling, Y. Wang, W. Dai, J. Yu, P. Liang, and D. Kong, "MTANet: multi-task attention network for automatic medical image segmentation and classification," *IEEE Trans. Med. Imag.*, vol. 43, no. 2, pp. 674–685, Feb. 2024.

[2] J. Bai, R. Liu, H. Zhao, Z. Xiao, Z. Chen, W. Shi, Y. Xiong, and L. Jiao, "Hyperspectral image classification using geometric spatial–spectral feature integration: A class incremental learning approach," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5531215.

[3] C. Bass, M. D. Silva, C. Sudre, L. Z. J. Williams, H. S. Sousa, P.-D. Tudosiu, F. Alfaro-Almagro, S. P. Fitzgibbon, M. F. Glasser, S. M. Smith, and E. C. Robinson, "ICAM-reg: Interpretable classification and regression with feature attribution for mapping neurological phenotypes in individual scans," *IEEE Trans. Med. Imag.*, vol. 42, no. 4, pp. 959–970, Apr. 2023.

[4] S. Li, P. Kou, M. Ma, H. Yang, S. Huang, and Z. Yang, "Application of semi-supervised learning in image classification: Research on fusion of labeled and unlabeled data," *IEEE Access*, vol. 12, pp. 27331–27343, 2024.

[5] Y. Han, H. Zhu, L. Jiao, X. Yi, X. Li, B. Hou, W. Ma, and S. Wang, "SSMU-net: A style separation and mode unification network for multimodal remote sensing image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5407115.

[6] X. Li, Q. Song, J. Wu, R. Zhu, Z. Ma, and J.-H. Xue, "Locally-enriched cross-reconstruction for few-shot fine-grained image classification," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 33, pp. 7530–7540, 2023.

[7] G. Yue, P. Wei, Y. Liu, Y. Luo, J. Du, and T. Wang, "Automated endoscopic image classification via deep neural network with class imbalance loss," *IEEE Trans. Instrum. Meas.*, vol. 72, pp. 1–11, 2023.

[8] Z. Wu, C. Liu, J. Wen, Y. Xu, J. Yang, and X. Li, "Selecting high-quality proposals for weakly supervised object detection with bottom-up aggregated attention and phase-aware loss," *IEEE Trans. Image Process.*, vol. 32, pp. 682–693, 2023.

[9] J. Wu, C. Qin, and G. Feng, "SMDC-net: Saliency-guided multihead distribution calibration network for few-shot object detection on remote sensing images," *IEEE Geosci. Remote Sens. Lett.*, vol. 21, pp. 1–5, 2024.

[10] K. Zheng, Y. Dong, W. Xu, W. Tan, and P. Huang, "Auto learner of objects co-occurrence knowledge for object detection in remote sensing images," *IEEE Geosci. Remote Sens. Lett.*, vol. 20, pp. 1–5, 2023.

[11] S. Li, Z. Zhou, M. Zhao, J. Yang, W. Guo, Y. Lv, L. Kou, H. Wang, and Y. Gu, "A multitask benchmark dataset for satellite video: Object detection, tracking, and segmentation," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5611021.

[12] Y. Sun, K. Song, T. Zhou, G. Wei, Z. Cheng, and C. Zhu, "A shared method of metal object detection and living object detection based on the quality factor of detection coils for electric vehicle wireless charging," *IEEE Trans. Instrum. Meas.*, vol. 72, pp. 1–17, 2023.

[13] P. Ni, X. Li, D. Kong, and X. Yin, "Scene-adaptive 3D semantic segmentation based on multi-level boundary-semantic-enhancement for intelligent vehicles," *IEEE Trans. Intell. Vehicles*, vol. 9, pp. 1722–1732, 2024.

[14] Y. Li, T. Shi, Y. Zhang, and J. Ma, "SPGAN-DA: Semantic-preserved generative adversarial network for domain adaptive remote sensing image semantic segmentation," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5406717.

[15] Y. Wang, G. Li, and Z. Liu, "SGFNet: Semantic-guided fusion network for RGB-thermal semantic segmentation," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 33, no. 12, pp. 7737–7748, Dec. 2023.

[16] A. Zhang, S. Li, J. Wu, S. Li, and B. Zhang, "Exploring semantic information extraction from different data forms in 3D point cloud semantic segmentation," *IEEE Access*, vol. 11, pp. 61929–61949, 2023.

[17] L. Wu, L. Fang, X. He, M. He, J. Ma, and Z. Zhong, "Querying labeled for unlabeled: Cross-image semantic consistency guided semi-supervised semantic segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, pp. 8827–8844, 2023.

[18] G. Li, Y. Wang, Z. Liu, X. Zhang, and D. Zeng, "RGB-T semantic segmentation with location, activation, and sharpening," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 33, no. 3, pp. 1223–1235, Mar. 2023.

[19] S. Fan, F. Zhu, Z. Feng, Y. Lv, M. Song, and F.-Y. Wang, "Conservative-progressive collaborative learning for semi-supervised semantic segmentation," *IEEE Trans. Image Process.*, vol. 32, pp. 6183–6194, 2023.

[20] W. Zhou, H. Zhang, W. Yan, and W. Lin, "MMSMCNet: Modal memory sharing and morphological complementary networks for RGB-T urban scene semantic segmentation," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 33, pp. 7096–7108, 2023.

[21] S. Chen, T. Ye, Y. Liu, and E. Chen, "SnowFormer: Context interaction transformer with scale-awareness for single image desnowing," 2022, *arXiv:2208.09703*.

[22] J. Bossu, N. Hautière, and J.-P. Tarel, "Rain or snow detection in image sequences through use of a histogram of orientation of streaks," *Int. J. Comput. Vis.*, vol. 93, no. 3, pp. 348–367, Jul. 2011.

[23] Y. Wang, S. Liu, C. Chen, and B. Zeng, "A hierarchical approach for rain or snow removing in a single color image," *IEEE Trans. Image Process.*, vol. 26, no. 8, pp. 3936–3950, Aug. 2017.

[24] S.-C. Huang, D.-W. Jaw, B.-H. Chen, and S.-Y. Kuo, "Single image snow removal using sparse representation and particle swarm optimizer," *ACM Trans. Intell. Syst. Technol.*, vol. 11, no. 2, pp. 1–15, Apr. 2020.

[25] D.-W. Jaw, S.-C. Huang, and S.-Y. Kuo, "DesnowGAN: An efficient single image snow removal framework using cross-resolution lateral connection and GANs," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 31, no. 4, pp. 1342–1350, Apr. 2021.

[26] B. Cheng, J. Li, Y. Chen, S. Zhang, and T. Zeng, "Snow mask guided adaptive residual network for image snow removal," 2022, *arXiv:2207.04754*.

[27] W. T. Chen, H. Y. Fang, J. J. Ding, C. C. Tsai, and S. Y. Kuo, "JSTASR: Joint size and transparency-aware snow removal algorithm based on modified partial convolution and veiling effect removal," in *Computer Vision—ECCV 2020* (Lecture Notes in Computer Science), vol. 12366, A. Vedaldi, H. Bischof, T. Brox, J. M. Frahm, Eds., Cham, Switzerland: Springer, 2020.

[28] J.-H. Kim, J.-Y. Sim, and C.-S. Kim, "Video deraining and desnowing using temporal correlation and low-rank matrix completion," *IEEE Trans. Image Process.*, vol. 24, no. 9, pp. 2658–2670, Sep. 2015.

[29] B. Yang, Z. Jia, J. Yang, and N. K. Kasabov, "Video snow removal based on self-adaptation snow detection and patch-based Gaussian mixture model," *IEEE Access*, vol. 8, pp. 160188–160201, 2020.

[30] J. Tian, Z. Han, W. Ren, X. Chen, and Y. Tang, "Snowflake removal for videos via global and local low-rank decomposition," *IEEE Trans. Multimedia*, vol. 20, no. 10, pp. 2659–2669, Oct. 2018.

[31] K. Zhang, R. Li, Y. Yu, W. Luo, and C. Li, "Deep dense multi-scale network for snow removal using semantic and depth priors," *IEEE Trans. Image Process.*, vol. 30, pp. 7419–7431, 2021.

[32] Y.-F. Liu, D.-W. Jaw, S.-C. Huang, and J.-N. Hwang, "DesnowNet: Context-aware deep network for snow removal," *IEEE Trans. Image Process.*, vol. 27, no. 6, pp. 3064–3073, Jun. 2018.

[33] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, and L. Antiga, "Pytorch: An imperative style, high-performance deep learning library," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 32, 2019, pp. 1–9.

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

[35] D. Engin, A. Genç, and H. K. Ekenel, "Cycle-dehaze: Enhanced CycleGAN for single image dehazing," 2018, *arXiv:1805.05308*.

[36] W.-T. Chen, H.-Y. Fang, C.-L. Hsieh, C.-C. Tsai, I.-H. Chen, J.-J. Ding, and S.-Y. Kuo, "ALL snow removed: Single image desnowing algorithm using hierarchical dual-tree complex wavelet representation and contradict channel loss," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 4176–4185.

YUXUAN LI was born in Jining, Shandong, China.

He is currently an Intern with Beijing Institute of Technology.

LIN DAI was born in 1977. He received the Ph.D. degree. He is currently an Associate Professor with the School of Computer Science, Beijing Institute of Technology.