Cloud and snow segmentation via transformer-guided multi-stream feature integration

remote sensing

MDPI

Article

Cloud and Snow Segmentation via Transformer-Guided Multi-Stream Feature Integration

Kaisheng Yu [1], Kai Chen [1], Liguo Weng [1], Min Xia [1,*] and Shengyan Liu [2]

[1] Collaborative Innovation Center on Atmospheric Environment and Equipment Technology, B-DAT, Nanjing University of Information Science and Technology, Nanjing 210044, China; 202212490080@nuist.edu.cn (K.Y.); 20211249015@nuist.edu.cn (K.C.); 002311@nuist.edu.cn (L.W.)
[2] Department of Computer Science, University of Reading, Whiteknights, Reading RG6 6DH, UK; yq835099@student.reading.ac.uk
* Correspondence: xiamin@nuist.edu.cn

Highlights

What are the main findings?

- A novel dual-branch network integrating Transformer and CNN streams for cloud and snow segmentation.
- The model effectively fuses global contextual features and local spatial details to distinguish spectrally similar surfaces.

What is the implication of the main finding?

- Significantly improves boundary accuracy and robustness against noise in complex remote sensing scenes.
- Achieves state-of-the-art performance on CSWV and SPARCS datasets, enabling reli-able operational cloud and snow monitoring.

Abstract

Cloud and snow often share comparable visual and structural patterns in satellite observations, making their accurate discrimination and segmentation particularly challenging. To overcome this, we design an innovative Transformer-guided architecture with complementary feature-extraction capabilities. The encoder adopts a dual-path structure, integrating a Transformer Encoder Module (TEM) for capturing long-range semantic dependencies and a ResNet18-based convolutional branch for detailed spatial representation. A Feature-Enhancement Module (FEM) is introduced to promote bidirectional interaction and adaptive feature integration between the two pathways. To improve delineation of object boundaries, especially in visually complex areas, we embed a Deep Feature-Extraction Module (DFEM) at the deepest layer of the convolutional stream. This component refines channel-level information to highlight critical features and enhance edge clarity. Additionally, to address noise from intricate backgrounds and ambiguous cloud-snow transitions, we incorporate both a Transformer Fusion Module (TFM) and a Strip Pooling Auxiliary Module (SPAM) in the decoding phase. These modules collaboratively enhance structural recovery and improve robustness in segmentation. Extensive experiments on the CSWV and SPARCS datasets show that our method consistently outperforms state-of-the-art baselines, demonstrating its strong effectiveness and applicability in real-world cloud and snow-detection scenarios.

Keywords: transformer; cloud snow; semantic segmentation; multibranch; deep learning

Academic Editor: Filomena Romano

Received: 27 July 2025
Revised: 21 September 2025
Accepted: 24 September 2025
Published: 29 September 2025

Citation: Yu, K.; Chen, K.; Weng, L.; Xia, M.; Liu, S. Cloud and Snow Segmentation via Transformer-Guided Multi-Stream Feature Integration. *Remote Sens.* **2025**, *17*, 3329. https://doi.org/10.3390/rs17193329

1. Introduction

As artificial satellite technology progresses, satellite imagery is increasingly utilized for Earth observation. Annually, approximately 67% of the Earth's surface experiences cloud cover [1]. Over 30% of the region is affected by seasonal snow, with 10% being permanently snow-covered [2]. The presence of clouds and snow in satellite images significantly impacts the effectiveness of Earth observation. Due to the plateau area, every year there will be heavy snow to bring great losses to animal husbandry, and timely detection of snow cover can substantially reduce personnel and material losses caused by snowstorms. However, the spectral characteristics of the panchromatic band of cloud and snow have high similarity, presenting a technical challenge in satellite cloud and snow image recognition.

On one side, cloud and snow coverage complicates remote sensing by obstructing targets or altering surface reflectance, thereby affecting image interpretation and processing. On the other side, their highly similar visual and spectral properties make it difficult to distinguish between them, especially in cloud segmentation [3] and snow reflection estimation [4]. Hence, developing accurate cloud and snow-detection techniques is crucial for enhancing the reliability of remote sensing-based analysis.

Recent advances in deep learning have revolutionized remote sensing image analysis, achieving significant breakthroughs in cloud and snow detection [5–8]. Many models have been developed in recent years to address these challenges [9–11], pushing the field forward considerably. However, even state-of-the-art solutions have limitations.

Prior to the rise of deep learning, detection techniques were generally divided into three main categories: heuristic-based, temporal-reflectance comparison, and classical machine learning approaches. Heuristic or rule-based methods rely on spectral and thermal differences between land covers—such as clouds, snow, and vegetation—typically using thresholding strategies. Notable examples include the Automatic Cloud Cover Assessment (ACCA) [12], the Normalized Difference Snow Index (NDSI), Fmask [13], and snow decision tree algorithms [14]. These approaches, while straightforward, are primarily grounded in low-level spectral cues and exhibit strong dependence on shortwave infrared and thermal imaging [15], making them highly sensitive to sensor variability and reducing their cross-platform robustness [16].

Temporal-reflectance-based approaches (or multi-temporal methods) track radiometric changes over time to detect transient events like clouds or recent snowfall. Algorithms such as Tmask [17], CS [18], and ATSA [19] represent this class. Despite their temporal advantage, these techniques often face issues such as high dependency on dense time-series data, limited ability to identify persistent snow cover, and vulnerability to natural surface changes—factors that constrain their scalability.

Supervised machine learning strategies—including support vector machines [20], random forests [21], and Bayesian classifiers [22]—aim to learn discriminative features from labeled samples. Shallow neural networks [17] have also been applied. However, these models often struggle with feature complexity and fail to generalize well in challenging environments, falling short when compared to more recent, deep learning-based alternatives.

The introduction of deep learning has substantially improved detection accuracy, generalizability, and processing efficiency. In particular, Convolutional Neural Networks (CNNs) have become the standard for large-scale segmentation of clouds and snow [23]. Lightweight architectures like RS-Net [16] and streamlined variants of U-Net [24] are designed for fast inference while maintaining high accuracy. By utilizing U-Net [25] as a backbone and optimizing network depth, these models achieve reduced computational costs and parameter overheads, yet still significantly outperform traditional solutions such as Fmask in terms of precision and robustness.

Efforts to enhance prediction accuracy are primarily focused on refining existing modules, introducing auxiliary components, or redesigning the overall network architecture. One common strategy involves incorporating advanced modules into conventional frameworks. For instance, CloudNet [26] employs ResNet18 as its backbone and integrates an improved Atrous Spatial Pyramid Pooling (ASPP) module. This module, placed after the backbone feature extraction, captures multi-scale contextual features from the deepest layers, thereby refining the segmentation of cloud and shadow boundaries. The overall performance surpasses that of DeepLabv2 [27].

Another approach centers on rethinking the network design itself, especially in the encoder-decoder structure. CDUNet [28] enhances the traditional UNet by introducing a booster branch—comprising convolution and dropout layers—during training. This addition facilitates more effective loss computation and contributes to faster network convergence. In the decoding stage, rather than relying on the basic hierarchical fusion strategy used in UNet, the authors propose a novel feature fusion layer that simultaneously integrates three different feature maps. This design not only strengthens the extraction of fine-grained texture cues but also suppresses high-frequency noise. As a result, CDUNet yields more precise segmentation at object boundaries and offers stronger global contextual awareness, which improves its adaptability across different spatial environments. The model has demonstrated superior generalization on multiple satellite datasets.

Beyond these accuracy-driven strategies, some research also explores lightweight architectures to balance prediction precision and computational efficiency. For example, SGBNet [29] significantly reduces model complexity and enhances runtime performance. However, when applied to cloud and snow-segmentation tasks, it shows limitations in maintaining high segmentation accuracy.

Recent research efforts [30–32] have extended the use of Transformers—originally developed for natural language processing—to vision-related tasks by leveraging their ability to capture long-range dependencies. Transformer models [33] utilize a multi-head self-attention mechanism, enabling them to aggregate information across the entire input and emphasize salient regions. The essence of this mechanism lies in modeling the interrelations among all pixels, where each pixel participates in the computation but contributes to varying degrees. This enables the model to achieve a global perceptual field and prioritize contextually important features.

Building on this, the Vision Transformer (ViT) was initially proposed by Dosovitskiy et al. [34], which applies a pure Transformer architecture directly to sequences of image patches for classification tasks. This patch-based modeling approach has shown superior performance over conventional convolutional architectures in multi-class image classification scenarios. However, despite its success in classification, ViT exhibits limitations when directly applied to dense prediction tasks like semantic segmentation, where maintaining spatial detail and local context is crucial.

To extend Transformer models to dense vision problems like detection and segmentation, including object localization and semantic parsing, Wang et al. [35] introduced the Pyramid Vision Transformer (PVT). This approach adopts ViT as its foundation while incorporating a hierarchical design that progressively reduces the spatial resolution of feature maps, thereby decreasing both memory consumption and computational cost. This makes PVT particularly suitable for pixel-level prediction tasks. Similarly, Wu et al. [36] proposed the Convolutional Vision Transformer (CvT), which embeds convolutional operations into the Transformer framework to improve representational capacity and overall performance. Furthermore, Zamir et al. [37] presented Restormer, a Transformer-based model tailored for image restoration tasks such as denoising, motion deblurring, and defocus correction, leveraging long-range dependencies to enhance reconstruction quality.

Despite their success in general vision tasks, these models encounter considerable difficulties when applied to remote sensing scenarios involving cloud and snow segmentation. The main challenge lies in the frequent occlusion of surface details by cloud and snow layers, which leads to degraded image quality. Additionally, the interaction between these elements and background features introduces substantial noise. The primary shortcomings of current Transformer-based approaches in this domain include:

(1) Limited robustness to noise and complex surface features, often resulting in false positives;

(2) Ineffective detection of small, isolated targets, contributing to omissions;

(3) Coarse delineation at cloud and snow boundaries, which hinders precise edge segmentation and affects overall accuracy.

This study tackles the aforementioned challenges by introducing a Transformer-driven multi-branch architecture for the detection and segmentation of clouds and snow in remote sensing imagery. After extensive experimentation and refinement, we present an enhanced Transformer framework for both the encoder and decoder stages, termed the Transformer Encoder Module (TEM) and the Transformer Fusion Module (TFM). The core of our network integrates TEM with the ResNet18 [38] convolutional backbone. While the Transformer contributes global self-attention, context modeling, and strong generalization capabilities [39], the convolutional branch provides robustness to geometric transformations such as translation, scaling, and distortion [40,41]. By fusing these complementary advantages, our dual-branch design ensures more effective semantic representation and spatial detail extraction.

In this study, we focus on the specific spectral bands of the Landsat-8 satellite, which are essential for cloud and snow classification. Specifically, the visible bands—such as the Blue (Band 2), Green (Band 3), and Red (Band 4) bands, along with the Near-Infrared (NIR, Band 5)—are primarily used. These bands are crucial because the contrast between snow and clouds is most evident in the visible and near-infrared regions. The Blue and Red bands help distinguish clouds and snow from their surroundings, while the NIR band is particularly sensitive to snow, allowing for better separation of snow from other land cover types.

To further boost feature representation, a Feature-Enhancement Module (FEM) is positioned between the Transformer and convolutional pathways, enabling mutual guidance and adaptive information exchange. This architecture improves the model's ability to capture subtle and scattered cloud-snow structures. Furthermore, a Deep Feature-Extraction Module (DFEM) is integrated at the deepest convolutional layer to enhance channel-level representations. By emphasizing salient feature dimensions, DFEM improves the model's ability to capture abstract representations and leads to more precise delineation along cloud and snow contours.

In the decoder, the Transformer Fusion Module (TFM) and the Strip Pooling Auxiliary Module (SPAM) jointly process multi-scale features from both encoder branches. This collaborative decoding strategy facilitates the integration of high-level semantics with low-level spatial cues, enhancing resistance to background interference and reducing classification errors. Consequently, the network delivers clearer and more reliable segmentation, particularly in complex cloud-snow boundary regions.

2. Methodology

2.1. Backbone

To capture multi-scale features during encoding, we adopt a hybrid architecture that integrates both a Transformer-based branch and a convolutional branch, as illustrated in Figure 1. Convolutional neural networks (CNNs) excel at modeling local spatial correla-

tions by operating on localized receptive fields, which helps minimize parameter count, reduce overfitting risk, and enhance the model's ability to learn translation-invariant representations [42]. In contrast, the Transformer architecture is adept at capturing long-range dependencies through its self-attention mechanism, offering a more stable alternative to recurrent neural networks (RNNs) that are prone to gradient vanishing or explosion when processing extended sequences. By leveraging this mechanism, the model can dynamically prioritize informative input regions, thereby improving contextual understanding.

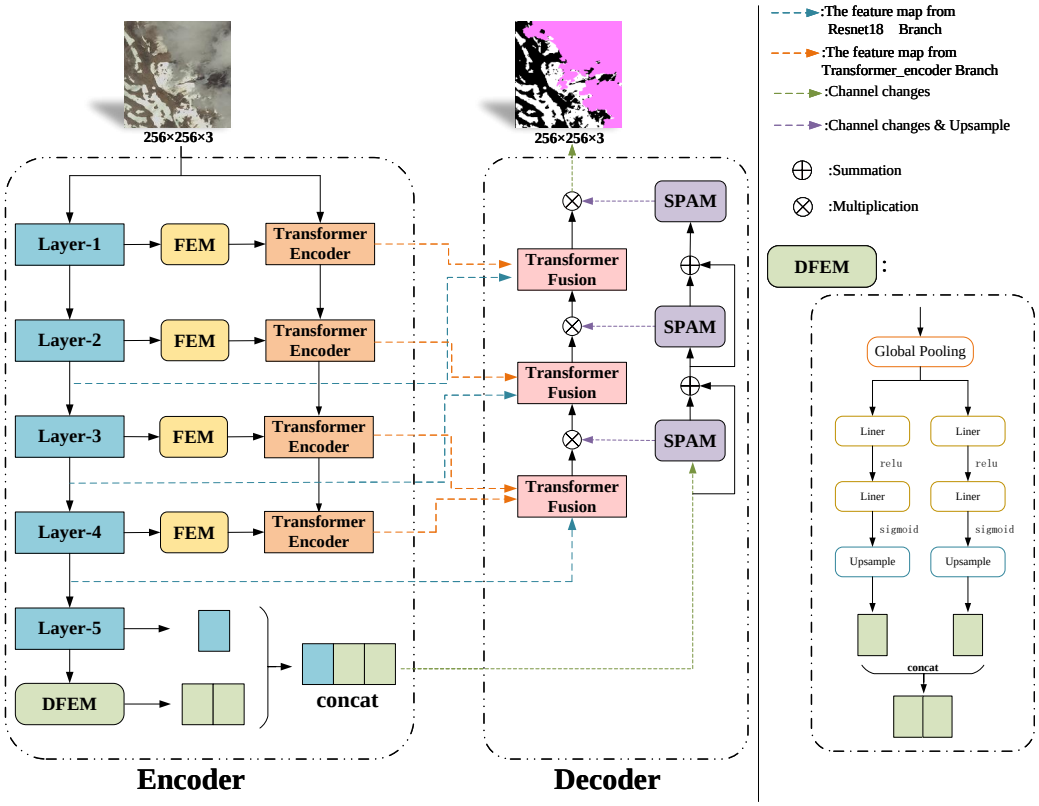

Figure 1. Multi-scale fusion attention network (Conv denotes convolutional layer, Avg indicates average pooling layer).

To take full advantage of these complementary strengths, we propose a Transformer Encoder Module (TEM), incorporated into the encoder stage. Its structure is shown in Figure 2a. By unifying local feature learning from CNNs with the global modeling capacity of Transformers, the proposed design achieves superior performance compared to architectures relying solely on either convolution or attention mechanisms.

The integration of CNNs and Transformers within the TEM allows the network to effectively capture both local spatial features and long-range dependencies. The convolutional branch excels in learning fine-grained spatial information, which is crucial for tasks requiring precise localization, such as segmenting objects with well-defined boundaries like clouds. In contrast, the Transformer branch leverages its global self-attention mechanism to model long-range dependencies, improving the network's ability to understand contextual relationships across the entire image. This hybrid design helps the model handle visually similar regions, such as snow and clouds, by combining detailed local features with broader contextual understanding, ultimately leading to more accurate segmentation in complex areas.

The dual-branch structure leverages the complementary advantages of both architectures, allowing for more effective extraction of spatial details and semantic context. To tackle the challenge posed by the visual similarity between clouds and snow—both

of which often share comparable shapes and spectral characteristics—we enhance the conventional ResNet18 convolutional pathway by integrating a Transformer Encoder Module (TEM). This addition significantly increases the network's ability to suppress mutual interference, thereby enhancing the precision of cloud and snow segmentation.

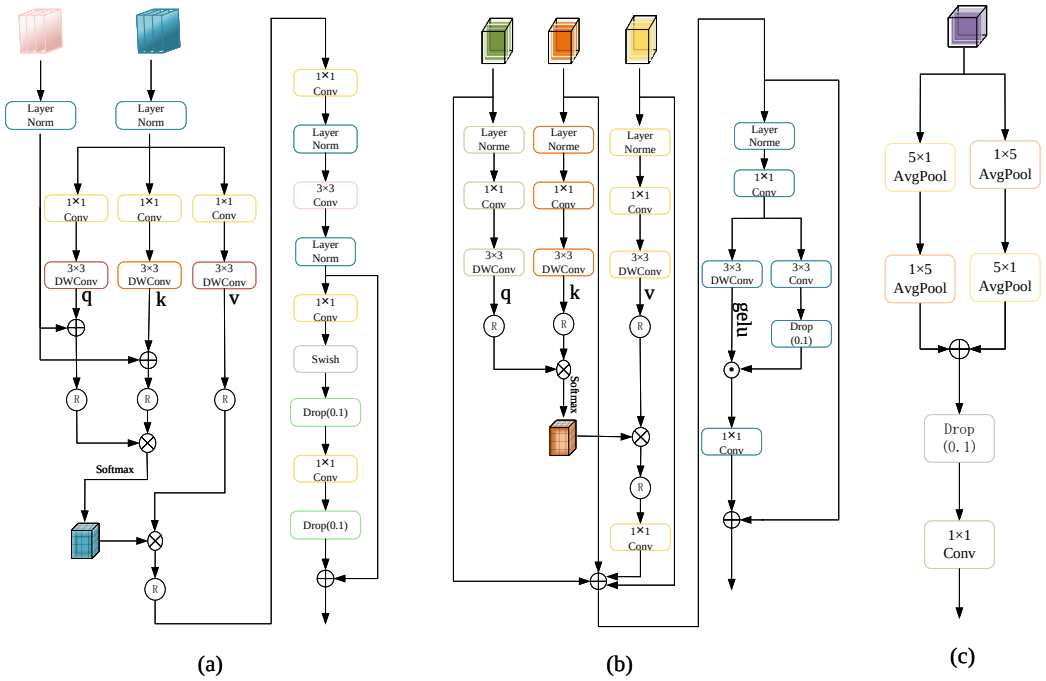

Figure 2. Block diagram of different stages. (**a**) TEM structure; (**b**) TFM structure; (**c**) SPAM structure. Conv denotes convolution layer. DWConv denotes depthwise separable convolution. Layer Norm denotes layer normalization. Gelu refers to the gelu activation function. Sigmoid refers to the sigmoid activation function. Softmax refers to the softmax function. Ⓡ indicates the Rearrange operation. ⊗ represents matrix multiplication. ⊙ represents element-wise multiplication. ⊕ indicates summation. Drop denotes the dropout operation.

While the Transformer-ResNet18 hybrid architecture we propose effectively captures both local and global dependencies, it is worth noting that alternative hybrid models, such as the Swin Transformer combined with CNNs, also have their own advantages. The Swin Transformer [40], for instance, adopts a shifted window mechanism, which is more computationally efficient and scalable compared to traditional Transformers. However, the Swin Transformer may not capture local spatial details as effectively as CNNs, especially in highly detailed regions like cloud and snow boundaries. In contrast, our approach benefits from the detailed feature extraction of ResNet18 and the global attention capabilities of the Transformer, offering a well-rounded solution that improves segmentation accuracy in challenging scenarios.

Details of the network architecture are provided in Table 1. Feature maps are organized into one to five stages according to their spatial dimensions. The proposed TEM incorporates multiple enhancements compared to conventional Transformer modules, especially in the Multi-Head Self-Attention (MHSA) and MLP components. Within the MHSA mechanism, a matrix fusion strategy is employed between the query (Q) and key (K) vectors to enhance the extraction of relevant image-level dependencies. This improves the model's ability to handle the complex dependencies between the cloud and snow regions, addressing their visual similarity more effectively.

Table 1. The architecture of the proposed network.

Level	Resnet18 Branch	Enhancemodule	TEM	DFEM	TFM	SPAM
L1	7×7 conv , 64	FEM	Transformer (d = 64 h = 8)		Transformer (d = 128 h = 4)	AvgPool Dr(0.1), 1×1
L2	3×3 max pool $\begin{bmatrix} 3 \times 3, 64 \\ 3 \times 3, 64 \end{bmatrix} \times 2$	FEM	Transformer (d = 64 h = 8)		Transformer (d = 64 h = 4)	AvgPool Dr(0.1), 1×1
L3	$\begin{bmatrix} 3 \times 3, 128 \\ 3 \times 3, 128 \end{bmatrix} \times 2$	FEM	Transformer (d = 128 h = 8)		Transformer (d = 64 h = 4)	AvgPool Dr(0.1), 1×1
L4	$\begin{bmatrix} 3 \times 3, 256 \\ 3 \times 3, 256 \end{bmatrix} \times 2$	FEM	Transformer (d = 256 h = 8)			
L5	$\begin{bmatrix} 3 \times 3, 512 \\ 3 \times 3, 512 \end{bmatrix} \times 2$	FEM		Avgpool liner Concate		

Meanwhile, the standard MLP component in Transformer architectures typically consists of a linear transformation followed by a non-linear activation function. We revise this component into a Convolutional Feedforward Perceptron (CFP), which incorporates 2D convolution operations with the Swish activation. Compared to traditional fully connected layers, the convolutional structure benefits from local connectivity and weight sharing, which reduces the overall number of trainable parameters and computational load. Additionally, this convolutional design allows the model to better preserve local spatial features during downsampling, further improving its ability to capture fine-grained details.

To further improve generalization, dropout is incorporated into the CFP. Specifically, a 1×1 convolution is used to encode inter-channel contextual information at the pixel level, and dropout with a probability of 0.1 is applied during training to randomly deactivate a portion of the neurons. This stochastic regularization helps reduce overfitting and marginally improves segmentation performance. These modifications make our model more resilient to noise and able to segment complex cloud and snow regions more accurately.

In summary, while alternative hybrid architectures such as the Swin Transformer + CNN may offer advantages in computational efficiency and scalability, the combination of ResNet18 and Transformer in our model provides a balanced solution that excels in both local and global feature extraction. The empirical results presented later demonstrate that our approach outperforms these alternatives in specific segmentation tasks, making it a more suitable choice for remote sensing applications that involve complex cloud and snow regions.

The mathematical formulation of the TEM module is as follows:

$$TEM_1 = Conv_{1\times1} MHSA(Norm(X_1), Norm(X_2)), \tag{1}$$

$$TEM_2 = Conv_{3\times3}(Norm(TEM_1)), \tag{2}$$

$$TEM_{\text{out}} = CFP(TEM_2), \tag{3}$$

In this context, X_1 and X_2 represent the inputs to the Transformer Encoder Module (TEM). The operation $Conv_{1\times1}$ refers to a 2D convolution with a 1×1 kernel, while $Conv_{3\times3}$ denotes a 2D convolution with a 3×3 kernel. MHSA stands for Multi-Head Self-Attention, Norm refers to layer normalization, and CFP represents the Convolutional Feedforward Perceptron.

The calculation process of MHSA is outlined as follows:

$$Q = DWC2D^{Q}_{3\times3}(C2D^{Q}_{(1\times1)}(Norm(X_1))), \tag{4}$$

$$K = DWC2D^{K}_{3\times3}(C2D^{K}_{(1\times1)}(Norm(X_1))), \tag{5}$$

$$V = DWC2D^{V}_{3\times3}(C2D^{V}_{(1\times1)}(Norm(X_1))), \tag{6}$$

$$Q^{'} = R(Q + Norm(X_2)), \tag{7}$$

$$K^{'} = R(K + Norm(X_2)), \tag{8}$$

$$MHSA_{\text{out}} = R(V' \cdot Softmax(K'Q'/\alpha)). \tag{9}$$

The calculation process of CFP is outlined as follows:

$$CFP_1 = Norm(TEM_2), \tag{10}$$

$$CFP_2 = Drop(Swish(C2D_{1\times1}(CFP_1))), \tag{11}$$

$$CFP_3 = Drop(C2D_{1\times1}(CFP_2)), \tag{12}$$

$$CFP_{\text{out}} = CFP_1 + CFP_3, \tag{13}$$

Here, $Q' \in \mathbb{R}^{HW\times C}$, $K' \in \mathbb{R}^{C\times HW}$, and $V' \in \mathbb{R}^{HW\times C}$ are derived from the original $\mathbb{R}^{H\times W\times C}$ tensor after reshaping. $C2D^{(\bullet)}_{1\times1}$ refers to a 2D convolution with a 1×1 kernel, while $C2D^{(\bullet)}_{3\times3}$ represents a depthwise separable 2D convolution with a 3×3 kernel. R denotes the rearrangement operation, and α is a learnable scaling factor that adjusts the pointwise product between K' and Q' before applying the softmax function. Softmax refers to the normalized exponential function. Swish represents the Swish activation function, and Drop indicates the dropout operation.

Accurate edge segmentation of clouds and snow remains a major challenge in target detection and segmentation tasks. To address this, we take advantage of the fact that the deepest layers of the backbone network contain a high number of channels, which capture rich contextual and semantic information. Accordingly, we introduce a Deep Feature-Extraction Module (DFEM) at the bottom of the backbone, as shown in Figure 1.

This module starts by compressing the spatial dimensions to produce a $1 \times 1 \times C$ representation, which summarizes each channel's global context through global average pooling, resulting in C global descriptors. These descriptors are then processed through two parallel transformation paths. In each path, a fully connected layer reduces the channel dimension, followed by a ReLU activation. The channel dimensions are then restored through another fully connected layer and refined with a sigmoid activation, helping the model focus on the most informative channels.

Next, spatial resolution is recovered, and to reinforce semantic and contextual cues—especially for delineating fine cloud and snow edges—the outputs of the two branches are merged by stacking features across channels. This fusion strategy encourages the network to focus more precisely on boundary localization. The above process can be formally described as:

$$y = G_{avg}(x), \tag{14}$$

$$y_1 = \text{Up}(\text{sigmoid}(\text{liner}(\text{relu}(\text{liner}(\text{y}))))), \tag{15}$$

$$y_2 = \text{Up}(\text{sigmoid}(\text{liner}(\text{relu}(\text{liner}(\text{y}))))), \tag{16}$$

$$\text{y}_{out} = CAT_1(y_1, y_2), \tag{17}$$

here, x and y_{out} denote the module's input and the output after feature mapping, respectively. G_{avg} refers to global average pooling, and CAT indicates concatenation along the channel dimension.

2.2. Feature-Enhancement Module (FEM)

Detecting and segmenting thin clouds and scattered small snow patches is challenging due to their wide dispersion and small size, making them prone to missed detections.

We believe this issue stems from insufficient fusion of location and category feature information. To improve accuracy, we weight the low-level features from the convolution branch alongside the high-level semantic features from the Transformer Encoder Module (TEM) branch. While the convolution branch preserves more spatial detail, the TEM branch captures higher-level features, making the convolution branch essential for mining deeper feature information and guiding the TEM branch with spatial context.

To enhance the low-level features, we introduced a Feature-Enhancement Module (FEM), as shown in Figure 3. This module strengthens spatial context, extracts multi-scale features, and highlights key elements. Consequently, it improves the model's ability to detect target boundaries and manage targets at various scales, boosting overall performance. The features from all three branches are concatenated and subsequently processed using depthwise separable convolution (dilation = 2), which restores channel capacity while broadening the effective receptive area, thereby improving the model's grasp of both local context and overall scene structure.

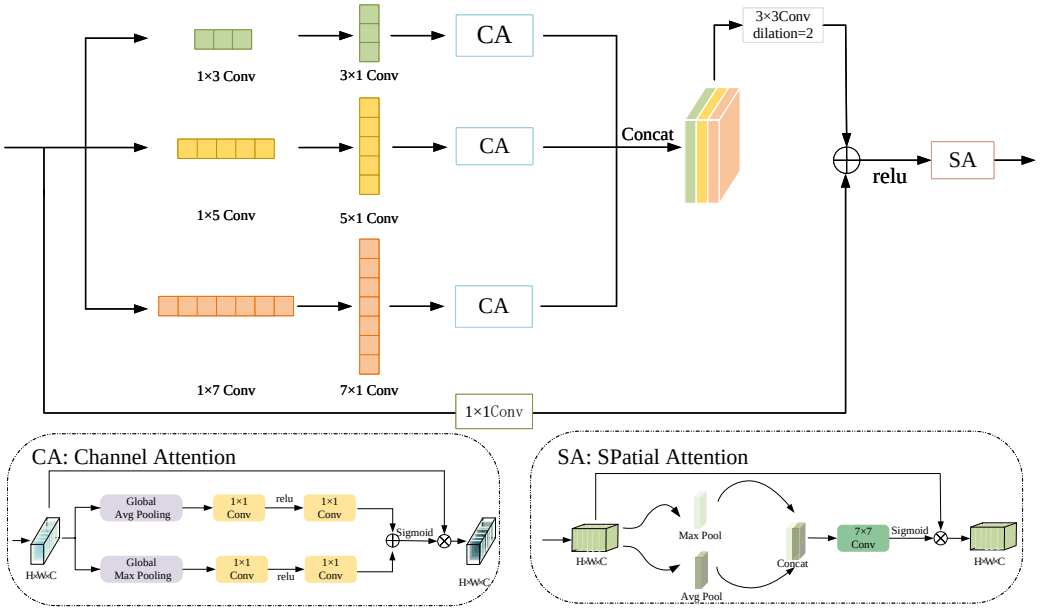

Figure 3. FEM structure diagram. Conv denotes convolution layer. Sigmoid refers to the sigmoid activation function. ReLU refers to the ReLU activation function. Rate indicates the dilation rate in dilated convolution.

Initial input weights for the FEM are adjusted via a 1×1 convolution. The output from the depthwise separable convolution is then added for feature fusion, followed by a ReLU activation function for nonlinear transformation and spatial attention. This allows the model to focus on specific spatial positions, improving the segmentation's spatial accuracy.

The channel attention module leverages both global max pooling and global average pooling to extract high-level features, enabling the capture of richer and more diverse semantic representations. A pointwise (1×1) convolution is subsequently employed to reduce the number of channels to one-sixteenth of the original, serving as a channel-wise feature selector that highlights the relative importance of each dimension. This operation is formulated in Equations (18) and (19):

$$\varphi_{max} = C2D_{1\times1}(G_{max}(x)), \tag{18}$$

$$\varphi_{avg} = C2D_{1\times1}(G_{avg}(x)), \tag{19}$$

where x denotes the input tensor. G_{max} and G_{avg} perform maximum and average pooling across spatial dimensions, respectively. $C2D_{1\times1}^{(\bullet)}$ applies a pointwise convolution for channel-wise transformation. After feature extraction, the ReLU activation function is applied to the feature map, helping suppress neuron activations without feedback, which enhances the model's sparse representation capability, noise resistance, and generalization. The channel dimensions are then restored through another 1×1 convolution. The weight vector from the global average pooling branch is added to the result from the global maximum pooling. The Sigmoid function is applied to recalibrate the feature map, which is subsequently combined with the original channel attention through element-wise multiplication. This operation is described by formula (20):

$$CA(x) = x \cdot \sigma(C2D_{1\times1}(\mathrm{Re}lu(\varphi_{avg}))) + C2D_{1\times1}(\mathrm{Re}lu(\varphi_{\max})), \tag{20}$$

where $CA(x)$ denotes the channel attention output, ReLU is the ReLU activation function, and σ refers to the Sigmoid activation function.

To enhance feature extraction, the spatial attention module integrates both maximum and average pooling. The resulting feature maps are then fused along the channel dimension through concatenation. After concatenation, a convolution with a 7×7 kernel reduces the number of channels from two to one. This large convolution kernel enables the extraction of a broader receptive field. The detailed computation process of the spatial attention module is presented in formula (21):

$$SA(x) = x \cdot \sigma(C2D_{7\times7}(CAT_1(MP(x), AP(x)))), \tag{21}$$

where $SA(x)$ denotes the spatial attention output, σ represents the Sigmoid activation function, $C2D_{7\times7}$ is a 2D convolution with a 7×7 kernel, and $\oplus$ indicates concatenation along the channel axis. MP and AP refer to maximum pooling and average pooling, respectively.

2.3. Transformer Fusion Module (TFM)

When a large area of snow and cloud overlap in the 2D image, cloud shadows may project onto the snow layer, creating significant color differences and interfering with surface elements in remote sensing images that resemble both clouds and snow. This results in incorrect attention to snow by the network, leading to misclassification. During decoding, a Transformer-driven fusion block is incorporated, illustrated in Figure 2b. This module effectively integrates the upsampled output from the decoder with multi-level feature data from the encoder branches. By leveraging diverse feature information, it strengthens the feature representation and enhances model performance. Additionally, it improves the network's ability to resist interference, particularly in regions where snow is covered by cloud shadows.

In this module, the weights output by the Multi-Head Self-Attention (MHSA) are passed through a convolutional embedding layer and combined with the three original weights input to the TFM. This integration allows the model to extract diverse feature information across different levels, optimizing the use of semantic details. By merging low-level and high-level features, the model achieves more comprehensive feature representations. In the MLP section, we replace the standard linear layer in most Transformers with 2D and depthwise separable convolutions, creating a Convolutional Feedforward Perceptron (CFP). This convolutional method extracts local patterns and spatial context via a sliding window, improving the network's capacity for fine-grained feature analysis. The use of

depthwise separable convolution reduces the number of parameters, improving training efficiency compared to traditional fully connected layers.

The detailed calculation process of the TFM is outlined as follows:

$$TFM_1 = C2D_{1\times1}(MHSA(Norm(Y_1), Norm(Y_2), Norm(Y_3))), \tag{22}$$

$$TEM_2 = TFM_1 + Y_1 + Y_2 + Y_3, \tag{23}$$

$$TEM_{\text{out}} = CFP(TEM_2), \tag{24}$$

where Y_1, Y_2, and Y_3 are the inputs to the TFM. $C2D_{1\times1}^{(\bullet)}$ denotes a 2D convolution with a 1×1 kernel. MHSA stands for Multi-head Self-Attention, Norm refers to layer normalization, and CFP represents Convolutional Feedforward Perceptron.

$$Q = DWC2D_{3\times3}^{Q}(C2D_{1\times1}^{Q}(Y_1)), \tag{25}$$

$$K = DWC2D_{3\times3}^{K}(C2D_{1\times1}^{K}(Y_2)), \tag{26}$$

$$V = DWC2D_{3\times3}^{V}(C2D_{1\times1}^{V}(Y_3)), \tag{27}$$

$$MHSA_{\text{out}} = R(V' \cdot Softmax(K' \cdot Q'/\alpha)). \tag{28}$$

The detailed computation process of the CFP is as follows:

$$CFP_1 = C2D_{1\times1}(Norm(TFM_2)), \tag{29}$$

$$CFP_2 = \delta(DWC2D_{3\times3}(CFP_1)), \tag{30}$$

$$CFP_3 = Drop(\text{Re}lu(BN(C2D_{3\times3}(CFP_1)))), \tag{31}$$

$$CFP_{out} = C2D_{1\times1}(CFP_2 \odot CFP_3) + TFM_2, \tag{32}$$

where $Q' \in \mathbb{R}^{HW\times C}$, $K' \in \mathbb{R}^{C\times HW}$, and $V' \in \mathbb{R}^{HW\times C}$ are obtained by reshaping the original tensor of size $\mathbb{R}^{H\times W\times C}$. R denotes the rearrangement operation, and α serves as a trainable scaling factor that controls the dot product magnitude between K' and Q' before the softmax. $C2D_{1\times1}^{(\bullet)}$ indicates a 2D convolution using a 1×1 kernel, while $C2D_{3\times3}^{(\bullet)}$ refers to a 2D depthwise separable convolution with a 3×3 kernel. δ is the GELU activation function, BN represents batch normalization, $\odot$ stands for element-wise multiplication, and Drop refers to the Dropout function.

2.4. Strip Pooling Auxiliary Module (SPAM)

Precisely distinguishing clouds and snow in satellite imagery is a challenge due to their similar colors and shapes. Existing methods often struggle to define precise boundaries, especially after down-sampling and up-sampling operations, which can result in the loss of fine details. To address these issues, we introduce the Strip Pooling Auxiliary Module (SPAM) in the decoding stage, as shown in Figure 2c.

SPAM consists of two parallel strip average pooling branches that extract spatial information from the feature map. These branches use convolution kernels of 1×5 and 5×1, respectively, to capture average values along the horizontal and vertical axes. This dual pooling mechanism enables the model to focus on both width and height dimensions, improving its ability to capture the shape and size of the target. The averaging process across both axes generates statistical features that better represent the spatial characteristics of the target, improving the smoothness and integrity of the feature map, and refining the segmentation of target boundaries.

After combining the outputs from both pooling branches, a dropout with a probability of 0.1 is applied. This regularization technique helps prevent overfitting by randomly eliminating neurons during the prediction phase, leading to more accurate segmentation results. The operations of SPAM are formally defined in formulas (33) and (34).

$$x' = Avg_{1\times5}(Avg_{5\times1}(x)) + Avg_{5\times1}(Avg_{1\times5}(x)), \tag{33}$$

$$y = C2D_{1\times1}(Drop(x')), \tag{34}$$

where x and y denote the input and output values of the module, respectively. $Avg_{5\times1}$ and $Avg_{1\times5}$ refer to average pooling layers with kernel sizes of 1×5 and 5×1, respectively. $C2D_{1\times1}(\cdot)$ indicates a 2D convolution with a 1×1 kernel. Drop refers to the Dropout operation.

3. Experiments

3.1. Dataset Introduction

(1) The Cloud and Snow (CSWV) Dataset [43] is used to evaluate the generalization performance of the proposed network. It consists of 27 high-resolution WorldView2 images, collected between June 2014 and July 2016 in the Cordillera Mountains, North America. The dataset features diverse landscapes, including forests, grasslands, ridges, and deserts, with cloud types (cirrus, cumulus, altocumulus, stratus) and snow forms (permanent, stable, discontinuous). These variations in shape, size, and texture make the dataset both comprehensive and challenging.

Each image is partitioned into 256 × 256 pixel patches, resulting in 3200 samples, which are divided into training (80%) and validation (20%) sets. To address the common issue of limited data in deep learning, augmentation techniques, such as translation, flipping, and random rotation, are applied, expanding the dataset to 10,240 training and 2560 validation images. Figure 4 displays sample images with cloud (pink), snow (white), and background (black) labels.

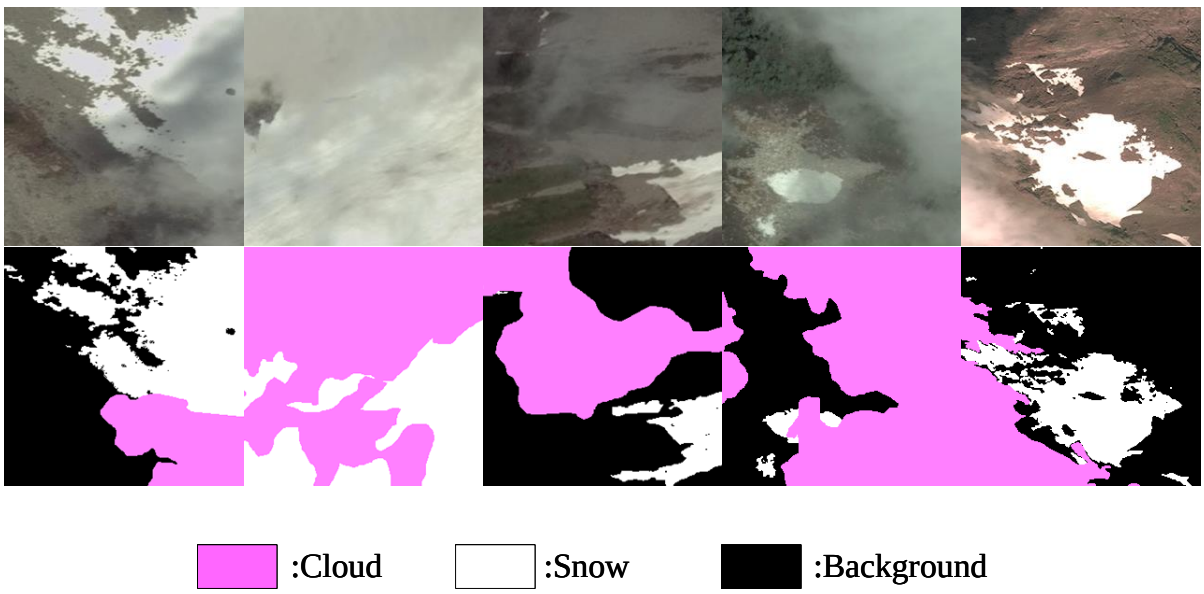

Figure 4. CSWV dataset part of the picture display.

Regarding the satellite data used in this study, we focus on specific spectral bands of the Landsat-8 satellite. Specifically, the visible bands—such as the Blue (Band 2), Green (Band 3), and Red (Band 4) bands, along with the Near-Infrared (NIR, Band 5)—are primarily used

for cloud and snow classification. These bands are crucial because the contrast between snow and clouds is most evident in the visible and near-infrared regions. The Blue and Red bands help distinguish clouds and snow from their surroundings, while the NIR band is particularly sensitive to snow, allowing for better separation of snow from other land cover types.

(2) The SPARCS Dataset [17] serves to evaluate the effectiveness of our method in multi-spectral image analysis. Developed by M. Joseph Hughes, it includes 80 Landsat-8 images sized 1000 × 1000 pixels, annotated with classes like clouds, cloud shadows, snow/ice, water, and background.

Due to GPU memory constraints, these images were cropped into smaller patches of 256 × 256 pixels, yielding 2000 samples. These were then divided into training and validation subsets with an 80:20 split. To enhance model generalization, data augmentation—incorporating translation, flipping, and rotation—expanded the training set to 6400 images and the validation set to 1600 images. Examples from the SPARCS dataset are presented in Figure 5, where cloud regions appear white, cloud shadows black, snow/ice light blue, water dark blue, and background gray.

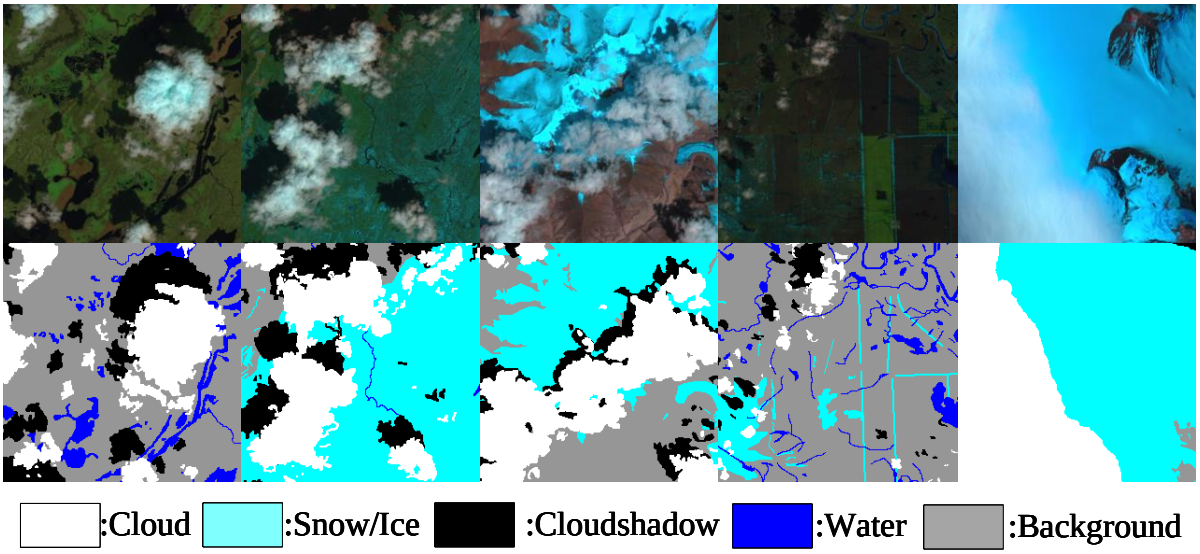

Figure 5. SPARCS dataset part of the picture display.

For the SPARCS dataset, we also use Landsat-8 spectral bands, specifically the visible bands (Blue, Green, Red) and the Near-Infrared (NIR) band. These bands are chosen because they effectively highlight the differences between snow and cloud regions, which are crucial for accurate classification in multi-spectral image analysis.

3.2. Experimental Details

We conducted experiments using PyTorch 2.2.2 with CUDA 12.1 support for GPU acceleration. [44]. The learning rate followed the "Steplr" schedule, computed as $\text{lrnew} = \text{lrinitial} \times \gamma^{\frac{epoch}{stepsize}}$. Initially set to 0.001, the learning rate decays by a factor of 0.98 every 3 epochs. We used the Adam optimizer [45], which is known for its stable and rapid convergence, with β_1 and β_2 set to 0.9 and 0.999, respectively. The experiments were performed on an NVIDIA GeForce RTX 3070 with 8 GB of memory, with a batch size of 4 due to GPU limitations. Training was carried out over 200 epochs. Performance on the CSWV and SPARCS datasets was assessed using key evaluation metrics: precision (P), recall (R), F1 score, pixel accuracy (PA), FWIoU, and MIoU. The corresponding formulas are listed as follows:

$$P = \frac{(TP)}{(TP) + (FP)}, \tag{35}$$

$$R = \frac{(TP)}{(TP) + (FN)}, \tag{36}$$

$$F_1 = 2 \times \frac{P \times R}{P + R}, \tag{37}$$

$$PA = \frac{\sum_{i=0}^{k} p_{i,i}}{\sum_{i=0}^{k} \sum_{j=0}^{k} p_{i,j}}, \tag{38}$$

$$FwioU = \frac{1}{\sum_{i=0}^{k} \sum_{j=0}^{k} p_{ij}} \sum_{i=0}^{k} \frac{\sum_{j=0}^{k} p_{ij} p_{ii}}{\sum_{0}^{k} p_{ij} + \sum_{j=0}^{k} p_{ji} - p_{ii}}, \tag{39}$$

$$MioU = \frac{1}{k+1} \sum_{i=0}^{k} \frac{p_{i,i}}{\sum_{j=0}^{k} p_{i,j} + \sum_{j=0}^{k} p_{j,i} - p_{i,i}}, \tag{40}$$

In these formulas, TP represents the correctly predicted cloud (or snow) pixels, while FP refers to incorrect predictions. FN indicates cloud (or snow) pixels that were misclassified. The number of categories, excluding the background, is denoted by k. For each category i, $p_{i,i}$ denotes the true positives, while $p_{i,j}$ represents pixels of category i predicted as category j.

3.3. Ablation Experiment

We conducted ablation studies on the CSWV cloud and snow dataset to evaluate the contribution of each module. Initially, we used the ResNet18 convolutional branch as the backbone, applying upsampling at each layer before connecting them for output. Then, we progressively added the modules (FEM, TEM, DFEM, SPAM, TFM) to assess their individual and collective impact. As shown in Table 2, the performance of each module was evaluated using MIoU, and the results demonstrate clear improvements with the inclusion of each module. To better visualize the effects of each module, we performed a thermal visualization experiment, which is illustrated in Figure 6.

FEM Ablation: To achieve precise localization and segmentation of thin clouds and small scattered snow patches, we designed the Feature-Enhancement Module (FEM) to facilitate cross-level connections between the two encoder branches. This module enhances the exchange of information and feature fusion. As shown in Table 2, the inclusion of FEM increased the network's MIoU to 86.33%, marking a 0.5% improvement. Thermal visualization in Figure 6d demonstrates that FEM improves the network's focus on cloud regions, enhances the detection of small, scattered targets, and reduces both missed and false detections.

TEM Ablation: Snow cover can interfere with cloud detection, reducing network attention to clouds and causing missegmentation. A single convolutional or transformer branch cannot fully extract the necessary features for accurate segmentation of both cloud and snow. To overcome this, we introduced the Transformer Encoder Module (TEM) as a parallel branch to the ResNet18 convolution branch, creating a dual-branch structure. This setup leverages the transformer's ability to capture long-range dependencies and the convolution's capability for extracting local details. The result is improved multi-scale feature extraction and better resistance to cloud-snow interference. As shown in

Table 2, incorporating TEM increased the MIoU to 87.52%, an improvement of 1.19%. Figure 6e illustrates that the TEM module significantly refines the focus on cloud prediction, improving segmentation accuracy.

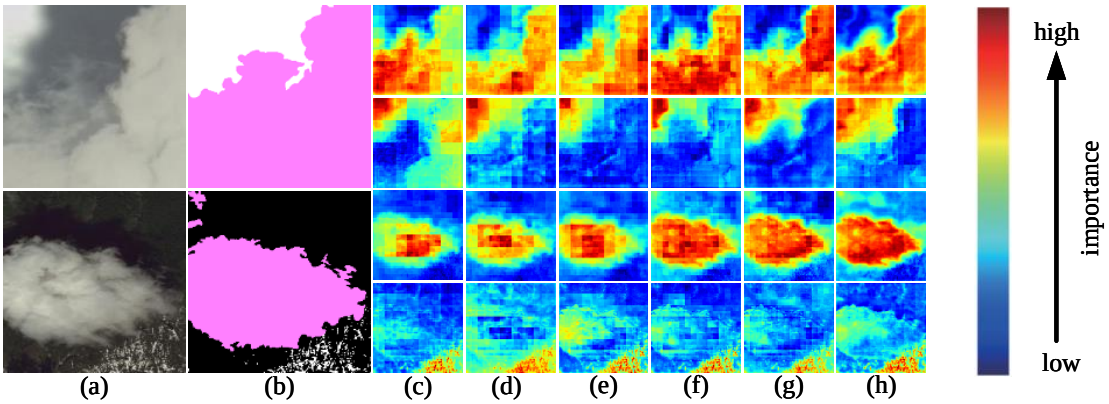

Figure 6. Heat map representation: (**a**) Real image, (**b**) Label, (**c**) ResNet18, (**d**) ResNet18 + FEM, (**e**) ResNet18 + FEM + TEM, (**f**) ResNet18 + FEM + TEM + DFEM, (**g**) ResNet18 + FEM + TEM + DFEM + SPAM, (**h**) ResNet18 + FEM + TEM + DFEM + SPAM + TFM. The first row shows cloud attention, and the second row shows snow attention.

Table 2. The performance of the network is evaluated progressively using the designed modules, with bold indicating the best-performing configuration and ↑ denoting performance improvement.

Method	MIoU (%)
Resnet18	85.83
Resnet18 + FEM	86.33 (↑ 0.5)
Resnet18 + FEM + TEM	87.52 (↑1.19)
Resnet18 + FEM + TEM + DFEM	87.64 (↑0.12)
Resnet18 + FEM + TEM + DFEM + SPAM	87.80 (↑0.16)
Resnet18 + FEM + TEM + DFEM + SPAM + TFM	**89.23** (↑1.43)

DFEM Ablation: The Deep Feature-Extraction Module (DFEM) was introduced at the base of the encoder's convolution branch, which holds the largest number of channels and contains rich semantic and contextual information. DFEM compresses and restores channels via linear layers, concatenates the output feature maps from the two parallel branches along the channel dimension, and maximizes the extraction of semantic and contextual information. This process enhances edge and texture details, improving the accuracy of edge segmentation for detection targets. As indicated in Table 2, the addition of DFEM raised the MIoU to 87.64%, a 0.12% increase. Heat map visualization in Figure 6f shows that the DFEM module helps the network focus on edge details, leading to more precise segmentation.

SPAM Ablation: The Strip Pooling Auxiliary Module (SPAM) was incorporated into the decoding stage to enhance the network's ability to perceive the shape, size, and boundaries of detected targets. This helps achieve precise segmentation of complex cloud and snow junctions. As shown in Table 2, adding SPAM increased the MIoU to 87.80%, a 0.16% improvement. Figure 6g highlights that SPAM enables the network to focus better on the cloud-snow junction, refining edge details and improving segmentation accuracy.

TFM Ablation: The Transformer Fusion Module (TFM) was designed in the decoding stage to fuse the feature information output by the upsampling decoding with the feature information extracted from the two encoder branches at different levels. This fusion process enhances feature mining, fully extracts spatial and semantic information, improves the

network's resistance to interference, and increases focus on the detection target. As seen in Table 2, the addition of TFM raised the MIoU to 89.23%, a 1.43% increase. Thermal visualizations in Figure 6h show that TFM significantly improves the network's attention to snow covered by cloud shadows in large-area snow images, reducing misjudgments and missed detections, while enhancing the robustness of the network.

3.4. Comparative Testing of Cloud and Snow (CSWV) Dataset

In this section, we compare our proposed network with several top-performing models, such as FCN, PAN, PSPNet, DeepLabV3Plus, BiSeNetV2, and others, to demonstrate its effectiveness. Each of these networks has distinct strengths. FCN uses a fully convolutional structure for pixel-wise classification. PSPNet captures multi-scale semantic information through pooling layers of different sizes, while DeepLabV3Plus incorporates an ASPP module with atrous convolutions at varying rates. BiSeNetV2, designed for real-time semantic segmentation, employs a dual-branch architecture to separately extract detailed and semantic features. In the Transformer-based models, PVT integrates feature pyramids with Transformers to leverage both methods' strengths, improving feature representation and small target detection. CvT enhances performance by introducing convolution operations within the Transformer framework. DBNet, a dual-branch model combining Transformer and convolutional networks, targets both semantic and spatial details to reduce false and missed detections in cloud-detection and -segmentation tasks.

Table 3 presents a comparison of various networks. For cloud detection, our network outperforms others in both recall (R) and F1 score, achieving 91.64% and 92.19%, respectively. Similarly, our network attains a recall of 93.59% and an F1 score of 94.25% for snow detection, surpassing other methods. While our network does not achieve the highest precision (P) in either cloud or snow detection, the gap compared to the top-performing method is minimal. Moreover, the proposed method achieves top performance in pixel accuracy (PA), frequency-weighted intersection over union (FWIoU), and mean intersection over union (MIoU), with values of 94.81%, 90.19%, and 89.23%, respectively. The findings confirm the outstanding capability and efficiency of our model.

Table 3. Comparison of network evaluation metrics on the cloud and snow (CSWV) dataset (bold indicates the best result).

Method	Cloud			Snow			Overall Results		
	P (%)	R (%)	F1 (%)	P (%)	R (%)	F1 (%)	PA (%)	FWIoU (%)	MIoU (%)
DFANet [46]	76.83	84.70	80.57	90.43	77.62	83.54	88.06	79.33	76.21
ESPNetV2 [47]	80.67	87.34	83.87	94.16	79.07	85.96	90.00	82.43	79.55
MFANet [48]	78.69	89.21	83.62	95.22	79.92	86.90	89.86	81.98	79.56
SGBNet [29]	79.59	89.33	84.18	95.32	78.38	86.03	90.14	82.62	79.82
BiSeNetV2 [49]	88.83	81.60	85.06	89.08	86.29	87.66	90.15	82.47	80.22
ENet [50]	80.80	88.81	84.61	94.35	80.20	86.70	90.45	83.07	80.43
PADANet [51]	82.10	87.64	84.78	93.36	81.20	86.86	90.50	83.15	80.53
DDRNet [52]	83.35	87.74	85.49	94.61	82.11	87.92	90.55	83.11	80.91
SP_CSANet [53]	84.57	87.31	85.92	94.08	83.87	88.68	91.34	84.46	82.13
DeepLabV3plus [54]	85.13	88.46	86.76	95.32	82.83	88.63	91.24	84.24	82.16
DenseASPP [55]	85.52	88.91	87.18	93.39	84.05	88.48	91.38	84.45	82.45
PVT [35]	85.85	87.04	86.44	92.60	86.09	89.23	91.50	84.65	82.57
MSPFANet [56]	85.14	88.46	86.76	95.96	84.06	89.61	91.72	85.03	82.96
MFENet [57]	86.13	88.73	87.41	93.31	85.34	89.15	91.82	85.18	83.18

Table 3. *Cont.*

Method	Cloud			Snow			Overall Results		
	P (%)	R (%)	F1 (%)	P (%)	R (%)	F1 (%)	PA (%)	FWIoU (%)	MIoU (%)
DABNet [58]	87.83	85.98	86.89	92.87	87.78	90.26	91.80	85.13	83.21
Restormer [37]	87.25	86.97	87.11	94.78	85.99	90.17	91.99	85.48	83.46
CvT [36]	89.42	86.17	87.77	93.03	87.53	90.20	92.12	85.69	83.79
CcNet [59]	91.77	86.81	89.22	95.48	87.27	91.19	92.31	85.91	84.56
LCDNet [60]	90.41	86.45	88.38	93.70	88.40	90.97	92.54	86.39	84.62
MCANet [61]	86.13	91.40	88.69	95.25	85.89	90.33	92.77	86.76	84.93
CDUNet [28]	87.42	90.03	88.71	95.11	87.97	91.40	92.92	86.96	85.31
ACFNet [62]	90.78	88.74	89.75	95.20	88.15	91.54	92.80	86.73	85.41
HRNet [63]	91.28	86.82	88.99	94.94	90.24	92.53	93.07	87.23	85.75
UNet [25]	91.75	87.57	89.61	**96.14**	88.62	92.23	93.04	87.17	85.80
SegNet [64]	91.84	88.05	89.90	95.70	88.58	92.00	93.10	87.27	85.91
DBNet [65]	91.89	87.28	89.52	94.00	90.22	92.07	93.31	87.66	86.10
CSDNet [43]	91.79	87.82	89.76	95.96	89.69	92.72	93.34	87.69	86.33
PSPNet [66]	**94.91**	87.44	91.02	90.99	94.12	92.53	93.78	88.40	87.22
DFN [67]	94.11	89.79	91.90	93.88	90.73	92.27	93.70	88.25	87.29
CloudNet [26]	93.67	88.58	91.05	94.51	91.79	93.13	93.97	88.76	87.57
PAN [68]	92.70	90.65	91.66	95.17	91.05	93.07	93.99	88.76	87.77
FCN8s [69]	92.44	90.47	91.45	95.80	90.64	93.15	94.06	88.90	87.81
Our	92.73	**91.64**	**92.19**	94.91	**93.59**	**94.25**	**94.81**	**90.19**	**89.23**

Figure 7 presents examples from various representative scenarios. In the selected examples, we highlight the network segmentation results at the same location with yellow boxes for easy comparison. In scenes with forest, grassland, and desert backgrounds, FCN8s, PAN, PSPNet, DeepLabV3Plus, and BiSeNetV2 miss or incorrectly detect scattered small clouds and snow. In contrast, our network accurately detects and segments nearly all clouds and snow in the image. In the third row, featuring urban backgrounds, our network's TFM module, with its self-attention mechanism, effectively fuses and decodes feature information. It extracts semantic details from supplementary hierarchical context, minimizing the impact of interference factors and enhancing segmentation accuracy and robustness. PAN also performs well by constructing a feature pyramid that captures multi-scale semantic information, improving robustness to size and position changes. However, only our network and PAN avoid misjudging snow caused by the white roof in the image, while other networks make errors. In the fourth and fifth rows, our network shows higher segmentation accuracy, especially at the irregular junctions of cloud and snow. In the sixth row, when large areas of snow and clouds overlap and cloud shadows are cast onto the snow, our network more accurately segments the snow beneath the cloud shadow. Finally, in the seventh and eighth rows, our network, aided by the DFEM module, provides more detailed edge segmentation of clouds and snow compared to other networks. These results demonstrate the superior performance and robustness of our proposed network across various backgrounds.

Figure 7. Comparison of segmentation performance across different networks under various environmental conditions in the CSWV dataset. (**a**) Real image, (**b**) ground truth label, (**c**) segmentation results of our network, (**d**) FCN8s results, (**e**) PAN results, (**f**) PSPNet results, (**g**) DeepLabV3+ results, (**h**) BiSeNetV2 results.

In the cloud and snow-segmentation task, we observed a significant presence of thin clouds and scattered small cloud clusters and snow patches in remote sensing images. To address this, we selected relevant images and segmentation results for comparative analysis, as shown in Figure 8. The results indicate that networks such as BiSeNetV2, DeepLabV3Plus, and PSPNet struggle with many missed and false detections when detecting thin clouds and small snow patches. This is primarily due to the noise present on the cloud boundaries, which can confuse the model's decision-making process. These methods fail to extract sufficient semantic and spatial information, leading to poor performance on small-scale, scattered clouds and snow. In contrast, our network demonstrates superior detection accuracy for these types of targets. As illustrated in the third, fourth, and eighth rows, our network performs better in both cloud-detection accuracy and edge detail segmentation. The FEM module, by weighting low-level features from the convolution branch and combining them with high-level features rich in semantic information from the TEM branch, guides the network with location information. This feature fusion allows our network to more accurately predict and segment thin clouds and scattered small snow blocks.

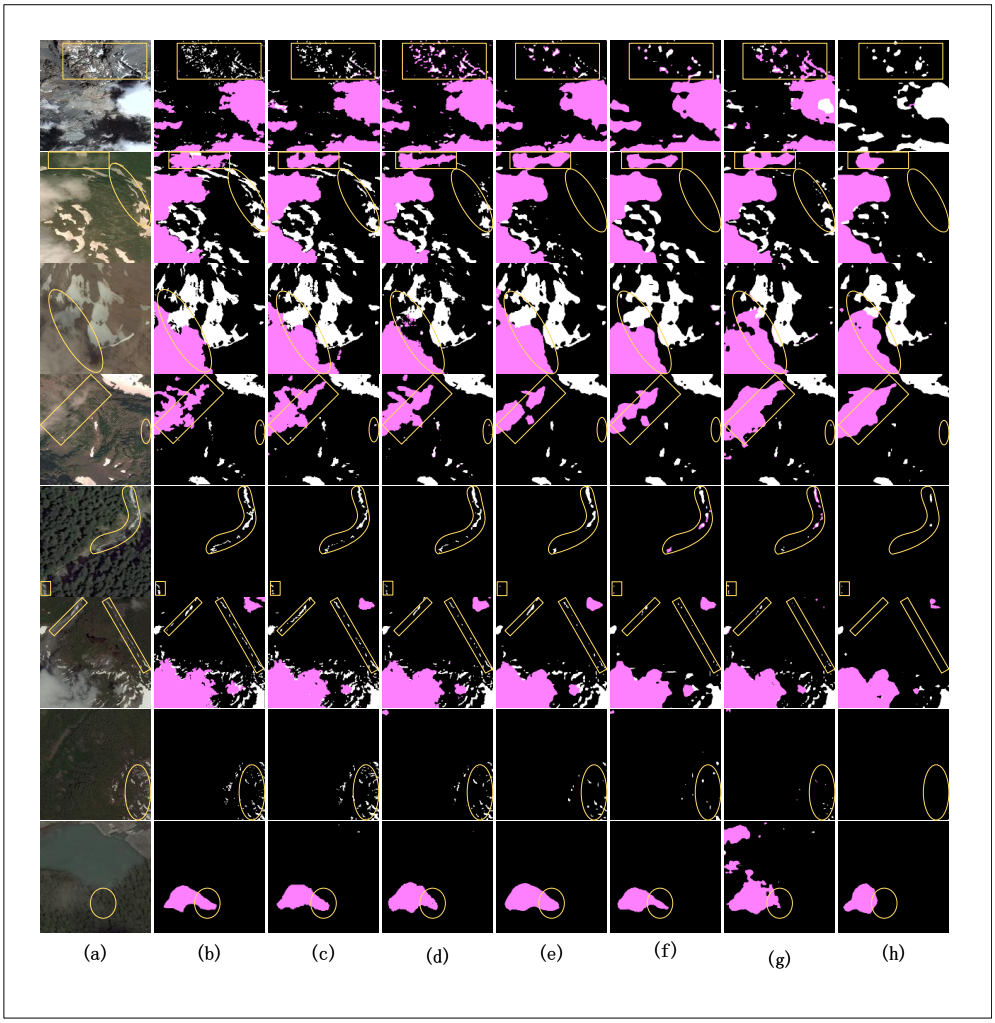

Figure 8. Comparison of segmentation effects of different networks for thin clouds, scattered small-sized cloud clusters and snow patches. (**a**) Real image, (**b**) label, (**c**) segmentation results of our net, (**d**) segmentation results of FCN8s, (**e**) segmentation results of PAN, (**f**) segmentation results of PSPNet, (**g**) segmentation results of DeepLabV3plus, (**h**) segmentation results of BiSeNetV2.

3.5. Generalization Experiment of SPARCS Dataset

To additionally assess our method's segmentation capabilities on multi-spectral satellite imagery, we performed generalization tests with the SPARCS dataset. The results are presented in Table 4. On the left side of the table, we can observe that our network achieves the highest F1 scores for snow/ice, water, and land categories, reaching 94.07%, 91.22%, and 95.72%, respectively. Although our F1 score for cloud and cloud shadow detection is not the highest, it is very close to the best-performing network. On the right side of the table, our network outperforms others in terms of PA, FWIoU, and MIoU, demonstrating its effectiveness and strong generalization ability.

Figure 9 presents the segmentation performance of various networks on the SPARCS dataset across different scenarios. The third, fourth, and seventh rows focus on the segmentation of scattered small-scale clouds and snow, while the fifth and sixth rows highlight the segmentation of thin clouds. From the results, it is clear that BiSeNetV2 and PSPNet struggle with large-scale misdetections, insufficient small target detection, and rough edge segmentation. While FCN8s performs better overall, it still has some error detections, particularly in segmenting the cloud-snow junction, where the details are lacking.

Table 4. Comparison of network performance across different evaluation metrics for the SPARCS dataset (bold indicates the best result).

Method	F1 (%)					Overall Results (%)		
	Cloud	CloudShadow	Snow/Ice	Water	Land	PA	FWIoU	MIoU
BiseNetV2	77.25	64.76	90.26	85.56	91.19	86.63	77.14	70.33
SegNet	83.00	61.03	91.14	83.42	92.75	88.50	80.45	71.33
SGBNet	84.43	64.47	89.35	84.09	92.59	88.61	80.14	72.02
ENet	84.30	66.61	89.91	85.89	93.12	89.19	81.06	73.37
PADANet	85.42	64.31	91.04	86.91	93.61	89.89	82.54	74.07
DeepLabV3plus	83.18	72.08	89.92	85.48	93.04	89.16	80.74	74.17
PSPNet	85.79	65.38	90.65	87.62	93.92	90.35	83.36	74.62
ESPNetV2	83.91	68.65	89.81	90.43	94.16	90.42	83.28	75.51
MSPFANet	87.00	66.09	92.12	88.69	94.56	91.02	84.33	76.22
DABNet	87.31	69.09	91.00	88.66	94.71	91.25	84.55	76.66
PAN	88.13	70.72	91.98	87.17	94.60	91.39	84.79	77.13
CvT	87.13	73.56	92.62	87.57	94.59	91.40	84.66	77.85
PVT	86.20	71.84	92.89	89.75	94.48	91.20	84.32	77.90
DBNet	88.79	72.71	92.91	88.30	95.14	92.19	86.23	78.70
SP_CSANet	89.43	74.23	92.45	88.26	95.23	92.35	86.33	79.15
UNet	88.31	74.88	92.94	88.90	95.06	92.14	85.94	79.27
CDUNet	87.12	76.00	92.79	90.30	95.33	92.27	86.19	79.68
CSDNet	88.95	**77.97**	92.76	87.49	95.23	92.38	86.23	79.83
FCN8s	**89.44**	75.41	91.79	90.12	95.57	92.75	87.00	79.96
Our	89.27	75.27	**94.07**	**91.22**	**95.72**	**93.02**	**87.33**	**81.49**

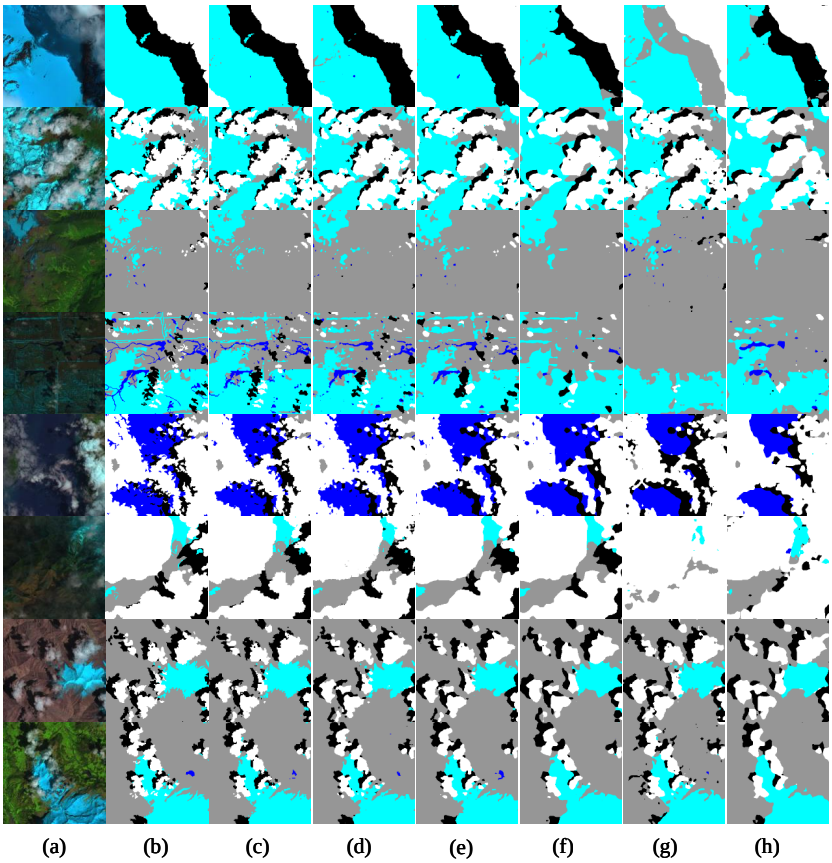

Figure 9. Comparison of segmentation results for different networks in various scenarios of the SPARCS dataset. (**a**) Real image, (**b**) label, (**c**) our network's segmentation, (**d**) FCN8s segmentation, (**e**) PAN segmentation, (**f**) PSPNet segmentation, (**g**) DeepLabV3plus segmentation, (**h**) BiSeNetV2 segmentation.

Our network, however, benefits from a dual-branch design that combines convolution and transformer branches in the encoding stage. By leveraging the strengths of both, we enhance the feature-extraction process and improve decoding. This significantly boosts the network's robustness and anti-interference capabilities, leading to more accurate segmentation. The FEM module, placed between the convolution and transformer branches, further strengthens cross-level information exchange, improving the network's ability to detect thin clouds and small targets.

The first and second rows demonstrate the network's ability to handle large snow and cloud areas, where cloud shadows are projected onto the snow layer in the remote sensing image. Our network effectively fuses multi-level feature information thanks to the TFM module, improving feature representation and segmentation accuracy. This results in better segmentation of clouds, snow, and cloud shadows, outperforming other networks. Additionally, the SPAM module extracts feature map averages in both horizontal and vertical directions, providing width and height information to enhance the network's ability to perceive the shape, size, and boundaries of the target. This contributes to a more accurate segmentation of complex junctions between cloud, snow, and cloud shadow. Compared to other networks, our approach demonstrates superior segmentation accuracy in these intricate regions.

These results confirm that our network outperforms others in the five-category multispectral remote sensing image-segmentation task, showcasing its effectiveness in complex semantic segmentation scenarios.

4. Conclusions

This paper presents a Transformer-based multi-branch feature fusion network designed for end-to-end cloud and snow segmentation in visible and multispectral high-resolution remote sensing images. The network integrates a transformer branch for extracting high-level semantic information and a convolution branch for capturing spatial location details. This fusion strengthens the network's robustness against cloud-snow interference and image noise, sharpening its attention to cloud detection.

The Feature-Enhancement Module (FEM) facilitates mutual guidance between the transformer and convolution branches during the encoding phase, promoting effective feature mining and fusion. This improves the network's attention to thin clouds, small scattered snow blocks, and clouds. To refine the segmentation boundaries, the Deep Feature-Extraction Module (DFEM) is introduced at the deepest layer of the convolution branch. It leverages fully connected layers to adjust the channels and extract deep-level contextual information, thereby enhancing boundary clarity.

To address background interference and rough segmentation of the cloud-snow junction, we design the Transformer Fusion Module (TFM) and Strip Pooling Auxiliary Module (SPAM) in the decoding stage. These modules boost the network's resilience to noise, enhance attention to snow detection, and improve the segmentation of irregular cloud-snow junctions.

Compared to existing methods, our approach significantly improves segmentation accuracy and handles complex scenarios effectively. Nevertheless, improvements can still be made. Upcoming efforts will aim to decrease the number of parameters in the model to boost inference speed without sacrificing segmentation accuracy.

Author Contributions: Conceptualization, K.Y., K.C. and L.W.; Methodology, K.Y., K.C. and L.W.; software, K.Y. and K.C.; validation, M.X. and S.L.; formal analysis, K.C. and K.Y.; investigation, M.X.; resources, M.X.; data curation, K.Y. and S.L.; writing—original draft preparation, K.Y. and L.W.; writing—review and editing, K.C.; visualization, K.Y.; supervision, L.W.; project administration,

M.X.; funding acquisition, M.X. All authors have read and agreed to the published version of the manuscript.

Funding: This work is supported in part by the National Natural Science Foundation of PR China (42075130).

Data Availability Statement: The data is available at 11 May 2025 https://pan.baidu.com/s/1or6bvqdaiZmxIpOsMbZBtg, accessed on 14 May 2025, extracted code c5wc. The code is publicly available at https://github.com/forever778/, accessed on 17 May 2025. Transformer-based-multi-branch-feature-fusion-network.

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

References

1. Zhao, L.; Chen, J.; Liao, Z.; Shi, F. Multi-Scale Feature Mixed Attention Network for Cloud and Snow Segmentation in Remote Sensing Images. *Remote Sens.* **2025**, *17*, 1872. [CrossRef]
2. Huang, K.; Sun, Z.; Xiong, Y.; Tu, L.; Yang, C.; Wang, H. Exploring Factors Affecting the Performance of Neural Network Algorithm for Detecting Clouds, Snow, and Lakes in Sentinel-2 Images. *Remote Sens.* **2024**, *16*, 3162. [CrossRef]
3. Hu, Z.; Weng, L.; Xia, M.; Hu, K.; Lin, H. HyCloudX: A multibranch hybrid segmentation network with band fusion for cloud/shadow. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2024**, *17*, 6762–6778. [CrossRef]
4. Liu, Z.; Chen, H.; Li, W.; Chen, K.; Qi, Z.; Liu, C.; Zou, Z.; Shi, Z. Learning to detect cloud and snow in remote sensing images from noisy labels. In Proceedings of the IGARSS 2024–2024 IEEE International Geoscience and Remote Sensing Symposium, Athens, Greece, 7–12 July 2024; pp. 7338–7341.
5. Zhang, Y.; Ye, C.; Yang, R.; Li, K. Reconstructing snow cover under clouds and cloud shadows by combining sentinel-2 and landsat 8 images in a mountainous region. *Remote Sens.* **2024**, *16*, 188. [CrossRef]
6. Yang, J.; Li, W.; Chen, K.; Liu, Z.; Shi, Z.; Zou, Z. Weakly supervised adversarial training for remote sensing image cloud and snow detection. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2024**, *17*, 15206–15221. [CrossRef]
7. Wu, Y.; Shi, C.; Shen, R.; Gu, X.; Tie, R.; Ge, L.; Sun, S. Snow Detection in Gaofen-1 Multi-Spectral Images Based on Swin-Transformer and U-Shaped Dual-Branch Encoder Structure Network with Geographic Information. *Remote Sens.* **2024**, *16*, 3327. [CrossRef]
8. Zhang, J.; Li, Y.; Yang, X.; Jiang, R.; Zhang, L. RSAM-Seg: A SAM-Based Model with Prior Knowledge Integration for Remote Sensing Image Semantic Segmentation. *Remote Sens.* **2025**, *17*, 590. [CrossRef]
9. Gu, H.; Gu, G.; Liu, Y.; Lin, H.; Xu, Y. Multi-Branch Attention Fusion Network for Cloud and Cloud Shadow Segmentation. *Remote Sens.* **2024**, *16*, 2308. [CrossRef]
10. Demil, G.; Torabi Haghighi, A.; Klöve, B.; Oussalah, M. Advances in Image-Based Estimation of Snow Hydrology Parameters: A Systematic Literature Review. *EGUsphere* **2024**, *2024*, 1–34.
11. Dumont, Z.B.; Gascoin, S.; Inglada, J. Snow and cloud classification in historical SPOT images: An image emulation approach for training a deep learning model without reference data. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2024**, *17*, 5541–5552. [CrossRef]
12. Irish, R.R.; Barker, J.L.; Goward, S.N.; Arvidson, T. Characterization of the Landsat-7 ETM+ automated cloud-cover assessment (ACCA) algorithm. *Photogramm. Eng. Remote Sens.* **2006**, *72*, 1179–1188.
13. Zhu, Z.; Woodcock, C.E. Object-based cloud and cloud shadow detection in Landsat imagery. *Remote Sens. Environ.* **2012**, *118*, 83–94. [CrossRef]
14. Pan, J.; Jiang, L.; Zhang, L. Wet snow detection in the south of China by passive microwave remote sensing. In Proceedings of the 2012 IEEE International Geoscience and Remote Sensing Symposium, Munich, Germany, 22–27 July 2012; pp. 4863–4866.
15. Bian, J.; Li, A.; Jin, H.; Zhao, W.; Lei, G.; Huang, C. Multi-temporal cloud and snow detection algorithm for the HJ-1A/B CCD imagery of China. In Proceedings of the 2014 IEEE Geoscience and Remote Sensing Symposium, Quebec City, QC, Canada, 13–18 July 2014; pp. 501–504.
16. Jeppesen, J.H.; Jacobsen, R.H.; Inceoglu, F.; Toftegaard, T.S. A cloud detection algorithm for satellite imagery based on deep learning. *Remote Sens. Environ.* **2019**, *229*, 247–259.
17. Hughes, M.J.; Hayes, D.J. Automated detection of cloud and cloud shadow in single-date Landsat imagery using neural networks and spatial post-processing. *Remote Sens.* **2014**, *6*, 4907–4926. [CrossRef]
18. Li, X.; Shen, H.; Zhang, L.; Zhang, H.; Yuan, Q.; Yang, G. Recovering quantitative remote sensing products contaminated by thick clouds and shadows using multitemporal dictionary learning. *IEEE Trans. Geosci. Remote Sens.* **2014**, *52*, 7086–7098.
19. Zhu, X.; Helmer, E.H. An automatic method for screening clouds and cloud shadows in optical satellite image time series in cloudy regions. *Remote Sens. Environ.* **2018**, *214*, 135–153.

20. Lee, K.Y.; Lin, C.H. Cloud detection of optical satellite images using support vector machine. *Int. Arch. Photogramm. Remote Sens. Spat. Inf. Sci.* **2016**, *41*, 289–293.
21. Ghasemian, N.; Akhoondzadeh, M. Introducing two Random Forest based methods for cloud detection in remote sensing images. *Adv. Space Res.* **2018**, *62*, 288–303. [CrossRef]
22. Hollstein, A.; Segl, K.; Guanter, L.; Brell, M.; Enesco, M. Ready-to-use methods for the detection of clouds, cirrus, snow, shadow, water and clear sky pixels in Sentinel-2 MSI images. *Remote Sens.* **2016**, *8*, 666. [CrossRef]
23. Ji, H.; Xia, M.; Zhang, D.; Lin, H. Multi-Supervised Feature Fusion Attention Network for Clouds and Shadows Detection. *ISPRS Int. J. Geo-Inf.* **2023**, *12*, 247. [CrossRef]
24. Wieland, M.; Li, Y.; Martinis, S. Multi-sensor cloud and cloud shadow segmentation with a convolutional neural network. *Remote Sens. Environ.* **2019**, *230*, 111203.
25. Ronneberger, O.; Fischer, P.; Brox, T. U-net: Convolutional networks for biomedical image segmentation. In Proceedings of the Medical Image Computing and Computer-Assisted Intervention—MICCAI 2015: 18th International Conference, Munich, Germany, 5–9 October 2015; pp. 234–241.
26. Xia, M.; Wang, T.; Zhang, Y.; Liu, J.; Xu, Y. Cloud/shadow segmentation based on global attention feature fusion residual network for remote sensing imagery. *Int. J. Remote Sens.* **2021**, *42*, 2022–2045. [CrossRef]
27. Chen, L.C.; Papandreou, G.; Kokkinos, I.; Murphy, K.; Yuille, A.L. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *40*, 834–848.
28. Hu, K.; Zhang, D.; Xia, M. CDUNet: Cloud detection UNet for remote sensing imagery. *Remote Sens.* **2021**, *13*, 4533. [CrossRef]
29. Pang, K.; Weng, L.; Zhang, Y.; Liu, J.; Lin, H.; Xia, M. SGBNet: An ultra light-weight network for real-time semantic segmentation of land cover. *Int. J. Remote Sens.* **2022**, *43*, 5917–5939.
30. Carion, N.; Massa, F.; Synnaeve, G.; Usunier, N.; Kirillov, A.; Zagoruyko, S. End-to-end object detection with transformers. In *Computer Vision—ECCV 2020*; Springer: Cham, Switzerland, 2020; pp. 213–229.
31. Zhao, J.; Jiao, L.; Wang, C.; Liu, X.; Liu, F.; Li, L.; Ma, M.; Yang, S. Knowledge Guided Evolutionary Transformer for Remote Sensing Scene Classification. *IEEE Trans. Circuits Syst. Video Technol.* **2024**, *34*, 10368–10384. [CrossRef]
32. Song, L.; Xia, M.; Xu, Y.; Weng, L.; Hu, K.; Lin, H.; Qian, M. Multi-granularity siamese transformer-based change detection in remote sensing imagery. *Eng. Appl. Artif. Intell.* **2024**, *136*, 108960.
33. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. *Adv. Neural Inf. Process. Syst.* **2017**, *30*.
34. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv* **2020**, arXiv:2010.11929.
35. Wang, W.; Xie, E.; Li, X.; Fan, D.P.; Song, K.; Liang, D.; Lu, T.; Luo, P.; Shao, L. Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, QC, Canada, 10–17 October 2021; pp. 568–578.
36. Wu, H.; Xiao, B.; Codella, N.; Liu, M.; Dai, X.; Yuan, L.; Zhang, L. Cvt: Introducing convolutions to vision transformers. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, QC, Canada, 10–17 October 2021; pp. 22–31.
37. Zamir, S.W.; Arora, A.; Khan, S.; Hayat, M.; Khan, F.S.; Yang, M.H. Restormer: Efficient transformer for high-resolution image restoration. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 5728–5739.
38. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
39. Zhu, T.; Zhao, Z.; Xia, M.; Huang, J.; Weng, L.; Hu, K.; Lin, H.; Zhao, W. FTA-Net: Frequency-Temporal-Aware Network for Remote Sensing Change Detection. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2025**, *18*, 3448–3460.
40. Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; Guo, B. Swin transformer: Hierarchical vision transformer using shifted windows. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021; pp. 10012–10022.
41. Yang, C.; Wang, J.; Meng, H.; Yang, S.; Feng, Z. Negative Class Guided Spatial Consistency Network for Sparsely Supervised Semantic Segmentation of Remote Sensing Images. *IEEE Trans. Circuits Syst. Video Technol.* **2025**, *35*, 657–669. [CrossRef]
42. Wu, J.; Fang, L.; Yue, J. TAKD: Target-Aware Knowledge Distillation for Remote Sensing Scene Classification. *IEEE Trans. Circuits Syst. Video Technol.* **2024**, *34*, 8188–8200. [CrossRef]
43. Zhang, G.; Gao, X.; Yang, Y.; Wang, M.; Ran, S. Controllably Deep Supervision and Multi-Scale Feature Fusion Network for Cloud and Snow Detection Based on Medium- and High-Resolution Imagery Dataset. *Remote Sens.* **2021**, *13*, 4805. [CrossRef]
44. Paszke, A.; Gross, S.; Chintala, S.; Chanan, G.; Yang, E.; DeVito, Z.; Lin, Z.; Desmaison, A.; Antiga, L.; Lerer, A. Automatic Differentiation in Pytorch. 2017. Available online: https://openreview.net/forum?id=BJJsrmfCZ (accessed on 15 March 2025).
45. Kingma, D.P.; Ba, J. Adam: A method for stochastic optimization. *arXiv* **2014**, arXiv:1412.6980.

46. Li, H.; Xiong, P.; Fan, H.; Sun, J. Dfanet: Deep feature aggregation for real-time semantic segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 9522–9531.
47. Mehta, S.; Rastegari, M.; Shapiro, L.; Hajishirzi, H. Espnetv2: A light-weight, power efficient, and general purpose convolutional neural network. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 9190–9200.
48. Hu, K.; Li, M.; Xia, M.; Lin, H. Multi-scale feature aggregation network for water area segmentation. *Remote Sens.* **2022**, *14*, 206.
49. Yu, C.; Gao, C.; Wang, J.; Yu, G.; Shen, C.; Sang, N. Bisenet v2: Bilateral network with guided aggregation for real-time semantic segmentation. *Int. J. Comput. Vis.* **2021**, *129*, 3051–3068.
50. Paszke, A.; Chaurasia, A.; Kim, S.; Culurciello, E. Enet: A deep neural network architecture for real-time semantic segmentation. *arXiv* **2016**, arXiv:1606.02147. [CrossRef]
51. Xia, M.; Qu, Y.; Lin, H. PANDA: Parallel asymmetric network with double attention for cloud and its shadow detection. *J. Appl. Remote Sens.* **2021**, *15*, 046512.
52. Hong, Y.; Pan, H.; Sun, W.; Jia, Y. Deep dual-resolution networks for real-time and accurate semantic segmentation of road scenes. *arXiv* **2021**, arXiv:2101.06085.
53. Qu, Y.; Xia, M.; Zhang, Y. Strip pooling channel spatial attention network for the segmentation of cloud and cloud shadow. *Comput. Geosci.* **2021**, *157*, 104940. [CrossRef]
54. Chen, L.C.; Zhu, Y.; Papandreou, G.; Schroff, F.; Adam, H. Encoder-decoder with atrous separable convolution for semantic image segmentation. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 801–818.
55. Yang, M.; Yu, K.; Zhang, C.; Li, Z.; Yang, K. Denseaspp for semantic segmentation in street scenes. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 3684–3692.
56. Lu, C.; Xia, M.; Lin, H. Multi-scale strip pooling feature aggregation network for cloud and cloud shadow segmentation. *Neural Comput. Appl.* **2022**, *34*, 6149–6162. [CrossRef]
57. Miao, S.; Xia, M.; Qian, M.; Zhang, Y.; Liu, J.; Lin, H. Cloud/shadow segmentation based on multi-level feature enhanced network for remote sensing imagery. *Int. J. Remote Sens.* **2022**, *43*, 5940–5960. [CrossRef]
58. Li, G.; Yun, I.; Kim, J.; Kim, J. Dabnet: Depth-wise asymmetric bottleneck for real-time semantic segmentation. *arXiv* **2019**, arXiv:1907.11357.
59. Huang, Z.; Wang, X.; Huang, L.; Huang, C.; Wei, Y.; Liu, W. Ccnet: Criss-cross attention for semantic segmentation. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 603–612.
60. Hu, K.; Zhang, D.; Xia, M.; Qian, M.; Chen, B. LCDNet: Light-weighted cloud detection network for high-resolution remote sensing images. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2022**, *15*, 4809–4823. [CrossRef]
61. Hu, K.; Zhang, E.; Xia, M.; Weng, L.; Lin, H. Mcanet: A multi-branch network for cloud/snow segmentation in high-resolution remote sensing images. *Remote Sens.* **2023**, *15*, 1055.
62. Zhang, F.; Chen, Y.; Li, Z.; Hong, Z.; Liu, J.; Ma, F.; Han, J.; Ding, E. Acfnet: Attentional class feature network for semantic segmentation. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 6798–6807.
63. Sun, K.; Zhao, Y.; Jiang, B.; Cheng, T.; Xiao, B.; Liu, D.; Mu, Y.; Wang, X.; Liu, W.; Wang, J. High-resolution representations for labeling pixels and regions. *arXiv* **2019**, arXiv:1904.04514. [CrossRef]
64. Badrinarayanan, V.; Handa, A.; Cipolla, R. Segnet: A deep convolutional encoder-decoder architecture for robust semantic pixel-wise labelling. *arXiv* **2015**, arXiv:1505.07293.
65. Lu, C.; Xia, M.; Qian, M.; Chen, B. Dual-branch network for cloud and cloud shadow segmentation. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 1–12.
66. Zhao, H.; Shi, J.; Qi, X.; Wang, X.; Jia, J. Pyramid scene parsing network. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 2881–2890.
67. Yu, C.; Wang, J.; Peng, C.; Gao, C.; Yu, G.; Sang, N. Learning a discriminative feature network for semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 1857–1866.
68. Li, H.; Xiong, P.; An, J.; Wang, L. Pyramid attention network for semantic segmentation. *arXiv* **2018**, arXiv:1805.10180. [CrossRef]
69. Long, J.; Shelhamer, E.; Darrell, T. Fully convolutional networks for semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 3431–3440.