Scientific Research Publishing

Journal of Computer and Communications, 2025, 13(1), 1-14
https://www.scirp.org/journal/jcc
ISSN Online: 2327-5227
ISSN Print: 2327-5219

ECD-Net: An Effective Cloud Detection Network for Remote Sensing Images

Hui Gao*, **Xianjun Du**

College of Electrical and Information Engineering, Lanzhou University of Technology, Lanzhou, China
Email: *wxhyq001@gmail.com, xdu@lut.edu.cn

How to cite this paper: Gao, H. and Du, X.J. (2025) ECD-Net: An Effective Cloud Detection Network for Remote Sensing Images. *Journal of Computer and Communications*, **13**, 1-14.
https://doi.org/10.4236/jcc.2025.131001

Received: December 10, 2024
Accepted: January 14, 2025
Published: January 17, 2025

Abstract

Cloud detection is a critical preprocessing step in remote sensing image processing, as the presence of clouds significantly affects the accuracy of remote sensing data and limits its applicability across various domains. This study presents an enhanced cloud detection method based on the U-Net architecture, designed to address the challenges of multi-scale cloud features and long-range dependencies inherent in remote sensing imagery. A Multi-Scale Dilated Attention (MSDA) module is introduced to effectively integrate multi-scale information and model long-range dependencies across different scales, enhancing the model's ability to detect clouds of varying sizes. Additionally, a Multi-Head Self-Attention (MHSA) mechanism is incorporated to improve the model's capacity for capturing finer details, particularly in distinguishing thin clouds from surface features. A multi-path supervision mechanism is also devised to ensure the model learns cloud features at multiple scales, further boosting the accuracy and robustness of cloud mask generation. Experimental results demonstrate that the enhanced model achieves superior performance compared to other benchmarked methods in complex scenarios. It significantly improves cloud detection accuracy, highlighting its strong potential for practical applications in cloud detection tasks.

Keywords

Deep Learning, Remote Sensing, Cloud Detection, MSDA, MHSA

1. Introduction

Optical remote sensing technology is increasingly utilized in Earth science research, yet cloud cover continues to pose a significant challenge. It is estimated that around 66% of the Earth's surface is persistently covered by clouds [1], which

*Corresponding author.

severely affects the accuracy of remote sensing (RS) data and restricts the broader application of RS technology in various fields. As a result, the precise identification of clouds in RS images has become a critical task for improving image clarity and ensuring the reliability of the data.

As RS technology continues to advance, cloud detection techniques have seen substantial progress and development. From the early reliance on traditional image processing techniques, such as threshold-based classification methods [2] [3], with the emergence and rapid advancement of deep learning in recent years, cloud detection technology has experienced a revolutionary transformation. Early traditional algorithms had significant limitations in terms of processing capability and adaptability, struggling to handle complex cloud formations and variable environmental conditions. However, as deep learning advances rapidly, deep learning-based cloud detection methods [4]-[16] have gradually become the dominant approach. Cloud detection can essentially be viewed as an image semantic segmentation task, and as image segmentation techniques have continually improved, cloud detection methods have also made notable breakthroughs. Deep learning methods, by automatically learning features, have not only significantly improved the accuracy of detecting complex cloud formations but have also far surpassed traditional methods in terms of generalization ability and processing efficiency.

Recently, deep neural networks have shown impressive performance in image segmentation tasks because of their outstanding feature extraction abilities, leading to their widespread use in cloud detection within remote sensing imagery. Among the cloud detection methods based on the U-Net framework, several innovative approaches have made significant advancements. For example, Cloud-FCN [4] enhances the U-Net architecture by incorporating the Inception module, enabling multi-scale feature extraction, which significantly improves cloud detection performance and outperforms traditional machine learning methods and threshold-based techniques. RS-Net [11] optimizes the U-Net structure by adjusting the number of embedded channels, reducing computational complexity while maintaining similar performance, and demonstrating outstanding segmentation results. For cloud detection in RS thumbnails, CDNet [9] introduces edge refinement techniques and a feature pyramid structure, effectively improving cloud detection accuracy in low-resolution images. MSCFF, on the other hand, enhances cloud detection in high-resolution imagery through multi-scale feature fusion. CDNetV2 [10] further advances cloud detection by maintaining high accuracy even in complex cloud and snow coexistence scenarios, laying a solid foundation for the continued development of cloud detection technology. Boundary net [13] delves deeply into multi-scale cloud and cloud mask boundary refinement, combining multi-scale feature fusion modules and differentiable boundary refinement networks. Although the model is more complex, it offers significant advantages in improving segmentation accuracy. AMCD-Net [14], taking into account the variability and complexity of clouds, integrates multi-level features and various attention mechanisms based on RS-Net, enabling more precise cloud detection in complex scenarios.

The U-Net-based Convolutional Neural Network (CNN) approach has achieved certain success in local feature extraction for RS image cloud detection tasks. However, it still faces significant challenges in establishing long-range dependencies and capturing global information. These limitations hinder the model's ability to handle complex scenes, particularly in distinguishing clouds from the surface and identifying multi-scale cloud structures.

In recent years, various models based on Transformer and Mamba (multi-modal attention mechanisms) architectures have been proposed and successfully applied to image segmentation tasks. Notable examples include Swin-UNet [17], SegFormer [18], EfficientVIT [19], U-MixFormer [20], U-Mamba [21], and CM-UNet [22], all of which have shown superior performance, particularly in capturing global information and long-range dependencies, far surpassing traditional CNN models.

Despite the strong performance of these approaches in segmentation tasks, their application to RS image cloud detection still faces several challenges. Specifically, the diversity and complexity of cloud layers, especially the similarity between thin clouds and the ground surface, make cloud mask generation a particularly difficult task. Moreover, RS image cloud detection requires multi-scale feature extraction of cloud layers, and accurate recognition of cloud structures at different scales remains an urgent problem.

To address these challenges and improve model performance in RS image cloud detection, this study proposes an enhancement to the classic U-Net framework. The architecture of its modules is restructured, with a particular focus on the multi-scale features of clouds in RS imagery. A Multi-Scale Dilated Attention (MSDA) [23] module is introduced to effectively incorporate multi-scale information and model long-range dependencies across different scales, significantly improving the model's ability to recognize clouds at various scales. Additionally, a Multi-Head Self-Attention (MHSA) [24] [25] mechanism is incorporated into the lower-level semantic feature extraction process, enhancing the model's ability to capture finer details, particularly in distinguishing thin clouds from the ground surface.

Building on this, the study also proposes a multi-path supervision mechanism to comprehensively supervise the cloud mask generation process. This ensures that the model learns cloud features at different scales and produces more accurate cloud masks in the output. The multi-path supervision not only improves the model's adaptability to multi-scale cloud structures but also enhances its robustness in complex scenarios, thereby significantly improving the accuracy of distinguishing thin clouds from the surface.

In conclusion, the proposed method in this study, by redesigning the network structure and incorporating techniques such as multi-scale feature extraction, long-range dependency modeling, and multi-path supervision, provides a more effective solution for RS image cloud detection tasks. The approach demonstrates strong potential for practical application.

2. Method

In this section, ECD-Net is proposed to address the task of cloud detection in RS images across various complex scenarios. Based on the U-Net framework, ECD-Net incorporates a MSDA module to effectively capture multi-scale feature representations of different cloud structures. Meanwhile, the network leverages the advantages of the MHSA module during low-level feature learning to further extract high-level semantic information of clouds, thereby enhancing the model's ability to distinguish features. With this design, ECD-Net is capable of generating more accurate cloud masks.

2.1. Overall Architecture

The complete framework of ECD-Net is illustrated in Figure 1. It is designed based on the classic U-Net architecture, incorporating several enhancements to improve its performance. Specifically, ECD-Net retains the fundamental encoder-decoder structure of U-Net, while introducing additional modules to better capture spatial and contextual information. These modifications enable the network to achieve more precise feature extraction and segmentation results.

The encoder comprises five stages that progressively extract features while reducing spatial resolution and enriching semantic information. Stage 1 applies a 3 × 3 convolutional layer (Conv 3 × 3), followed by batch normalization (BN) and the ReLU activation function, to extract initial features while preserving the input's resolution. Stage 2 introduces downsampling with a Conv 3 × 3 (stride = 2) to capture deeper feature representations, complemented by additional convolutional and normalization layers for feature refinement. Stages 3 and 4 incorporate MSDA Modules, enabling the model to effectively capture multi-scale cloud features and extract comprehensive global semantics. In Stage 5, a MHSA Modules is employed to model global dependencies, extracting high-level semantic features that serve as the foundation for the decoder.

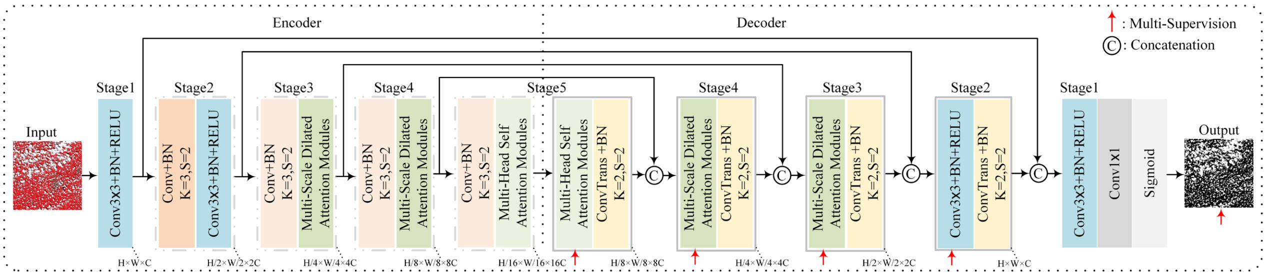

Figure 1. The architecture of the ECD-Net.

The decoder mirrors the encoder's structure but focuses on progressively recovering the spatial dimensions of the feature maps to generate the cloud mask.

Starting from the output of the encoder's final stage, Stage 5 of the decoder utilizes transposed convolution (ConvTrans + BN, K = 2, S = 2) for upsampling and incorporates a MHSA Modules to integrate global semantic information. Stages 4 and 3 continue upsampling through transposed convolution, leveraging MSDA Modules to enhance feature refinement. Skip connections are employed to fuse features from corresponding encoder stages, ensuring information completeness and preserving spatial details. Stage 2 further restores feature resolution closer to the original image size. Finally, Stage 1 applies a 1 × 1 convolutional layer to produce the cloud mask, which is normalized to a probability distribution in the range [0, 1] using a Sigmoid function.

The depth design of the attention module in both the decoder and encoder is the same [2, 2, 1].

2.2. MSDA Module and MHSA Module

This paper introduces the MSDA [23] module and the MHSA [24] [25] module to obtain multi-scale and global features. As represented in **Figure 2(a)**, the input first passes through Conditional Positional Encoding (CPE) [26] to incorporate positional information. This is followed by Layer Normalization (LayerNorm) to stabilize training. Then, the MSDA is applied to capture semantic information at different scales, with the output added back to the input via a residual connection to preserve the original information. Layer normalization is applied again, and an MLP is utilized to enhance the feature representation. In the end, another residual connection is applied to complete the module's output. The MHSA module (**Figure 2(b)**) follows a similar architecture to the MSDA module, leveraging the MHSA mechanism to model long-range dependencies within the input features.

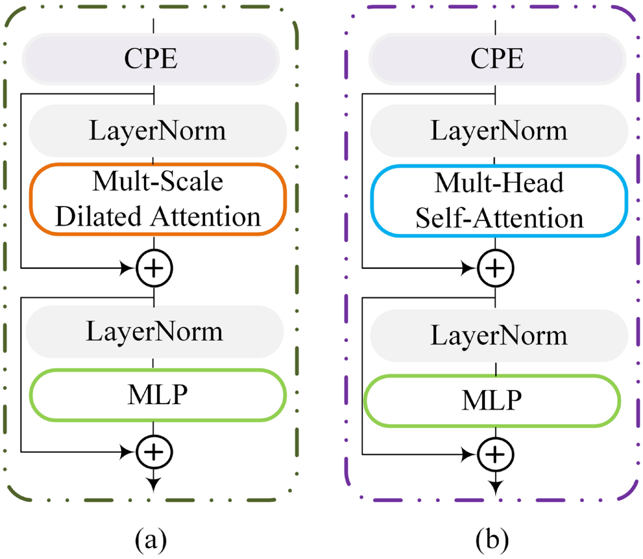

Figure 2. (a) MSDA module, (b)MHSA module.

As shown in **Figure 3**, the structure of MSDA changes such that the inputs Q, *K*, and *V* are generated through linear layers. Next, the channels are divided into *n* distinct attention heads., utilizing a SWDA (Sliding Window-based Sparse Attention) [23] mechanism to each head with varying dilation rates.

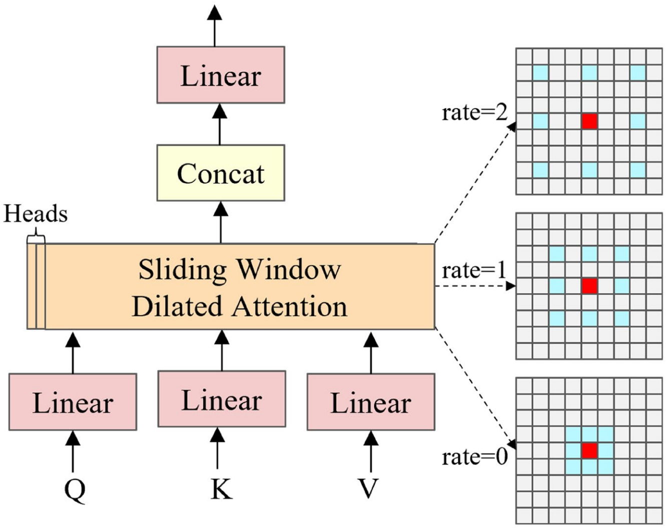

Figure 3. The structure of MSDA.

The operation of SWDA can be described as follows: given three matrices—*Q*, *K* and *V* as inputs, SWDA computes attention scores for each query vector q_{ij} at position (*i*, *j*) by selecting sparse *K* and *V* vectors within a sliding window centered at that point. The sparsity is controlled by the dilation rate *r*.

$$x_{ij} = \text{Attention}\left(q_{ij}, K_r, V_r\right) = \text{Softmax}\left(\frac{q_{ij}K_r^{\mathrm{T}}}{\sqrt{d_k}}\right)V_r \quad (1)$$

2.3. Loss Function

This paper adopts a multi-supervision approach to train ECD-Net, focusing on improving the model's performance. During training, the loss function uses BCE Loss [27], which calculates the pixel-wise binary classification error between the predicted results and the ground truth labels, effectively optimizing the model parameters. To further enhance the training performance, we design a multi-supervision loss function ($\mathcal{L}_{MS}$) based on BCE Loss. This function provides supervision at multiple-path, helping the model learn richer feature representations and improving the network's ability to adapt to complex scenarios.

Specifically, our $\mathcal{L}_{MS}$ computes the loss at each stage and combines them through a weighted sum, allowing the model to integrate feedback from different layers and avoid overly relying on features from any specific layer. With this multi-level and multi-scale supervision mechanism, the model can better learn both fine-grained details and global semantics in the image, further improving the accuracy and robustness of cloud detection.

The specific definition of this $\mathcal{L}_{MS}$ is as follows:

$$\mathcal{L}_{MS}\left(Y,\hat{Y}\right)=\sum_{n=1}^{5}\mathcal{L}_{BCE}\left(Y,\hat{Y}_n\right) \tag{2}$$

where the $\hat{Y}_n$ represents the cloud mask generated at each stage.

3. Dataset and Experimental Setup

3.1. Dataset

The performance of ECD-Net was evaluated using the GF1-WHU dataset [28] and the SPARCS dataset [29]. The GF1-WHU dataset includes 86 satellite images for training and 22 for testing, covering diverse cloud conditions and land types. To simplify the cloud detection task, cloud shadows in the cloud mask images were treated as background. RGB image patches with a resolution of 384 × 384 pixels were generated through cropping, resulting in 2012 training patches and 516 testing patches. During training, the dataset was randomly split into training and validation subsets in an 8:2 ratio. The SPARCS dataset consists of 80 RS images with a resolution of 1000 × 1000 pixels. These images were manually divided into 64 training images and 16 testing images based on different land surface types. The images were further cropped into patches of 384 × 384 pixels, yielding 576 patches for the training set and 144 patches for the testing set. During training, the training data was randomly split into training and validation subsets in a 9:1 ratio.

3.2. Implementation Details

In this study, the proposed ECD-Net was trained using the PyTorch framework [30] and optimized with the AdamW optimizer [31]. A cosine annealing scheduler with linear warm-up was employed to alter the learning rate. During training, the batch size was set to 24, with an initial learning rate of 4e-4 and a weight decay of 1e-3. The experiments ran for a total of 150 epochs. All experiments were conducted on a Windows 11 operating system and executed on an NVIDIA GeForce RTX 4090 GPU. This paper uses accuracy [14], Jaccard Index (Jaccard) [32] and F1-Score [33] to evaluate the proposed model.

4. Experiments

4.1. Results from the Quantitative Evaluation on the GF1-WHU Dataset

According to the quantitative comparison results presented in **Table 1** and **Figure 4**, ECD-Net excels in key evaluation metrics such as accuracy (97.37%), Jaccard index (87.09%), and F1 score (93.10%), demonstrating its high precision and superior segmentation performance in cloud detection tasks. Despite its higher computational cost (GFLOPs of 55.711), which requires more computational resources, its outstanding performance makes it the model of choice for tasks requiring high accuracy. In comparison, AMCD-Net, with a GFLOPs of 37.187, strikes a better balance between performance and computational efficiency. While it slightly trails

ECD-Net in accuracy and F1 score, it still delivers excellent results. U-Net, as a traditional network architecture, has a higher computational complexity (GFLOPs of 89.955), but still performs well regarding accuracy (96.92%) and F1 score (91.92%), making it a widely used baseline model for cloud detection tasks. RS-Net, with a reduced computational load (GFLOPs of 38.183), performs exceptionally well in optimizing computational efficiency. Its accuracy (96.98%) and F1 score (91.93%) are similar to U-Net's, making it a more efficient alternative.

Table 1. The results of different methods.

Method	Parmas (M)	GFLOPs	FPS	Accuracy (%)	Jaccard (%)	F1-Score (%)
U-Net	17.263	89.955	160.26	96.92	85.04	91.92
RS-Net	9.389	38.183	177.45	96.98	85.07	91.93
AMCD-Net	10.025	37.187	102.86	97.24	86.53	92.78
ECD-Net (ours)	13.558	55.711	88.58	97.37	87.09	93.10

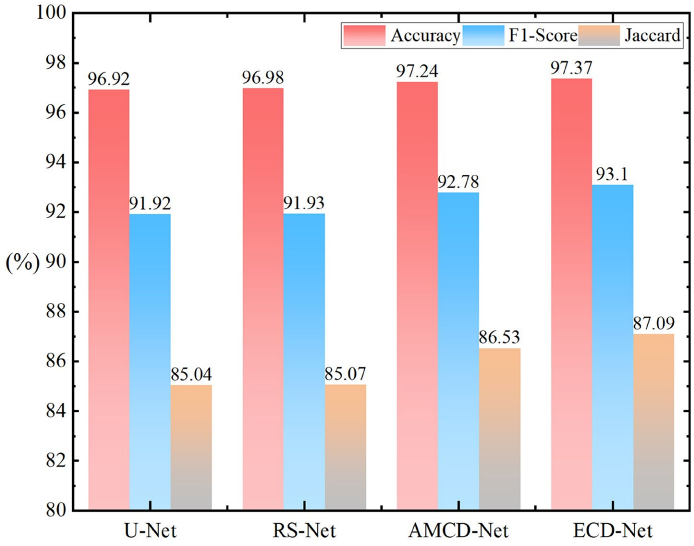

Figure 4. Performance of different methods.

However, ECD-Net's inference speed (88.58 FPS) is significantly lower compared to other methods. This can be attributed to the computationally intensive self-attention mechanism and the multi-path supervision strategy employed in the model. While these techniques effectively enhance detection accuracy, they also increase the computational burden, resulting in slower inference speed.

4.2. Visualization Results of Different Methods on the GF1-WHU Dataset

This paper compares the visualization results of four models across three different scenes, as shown in Figure 5. In the thin cloud regions over the water scene, all methods display inconsistent rates of missed and false detections. Among them, ECD-Net performs the most accurately, with clear boundaries between the cloud layers and the background, showing minimal missed and false detections. Although AMCD-Net is slightly inferior to ECD-Net, it still maintains high segmentation accuracy, with well-defined boundaries and fewer errors. In contrast, U-Net and RS-Net show larger areas of missed detections.

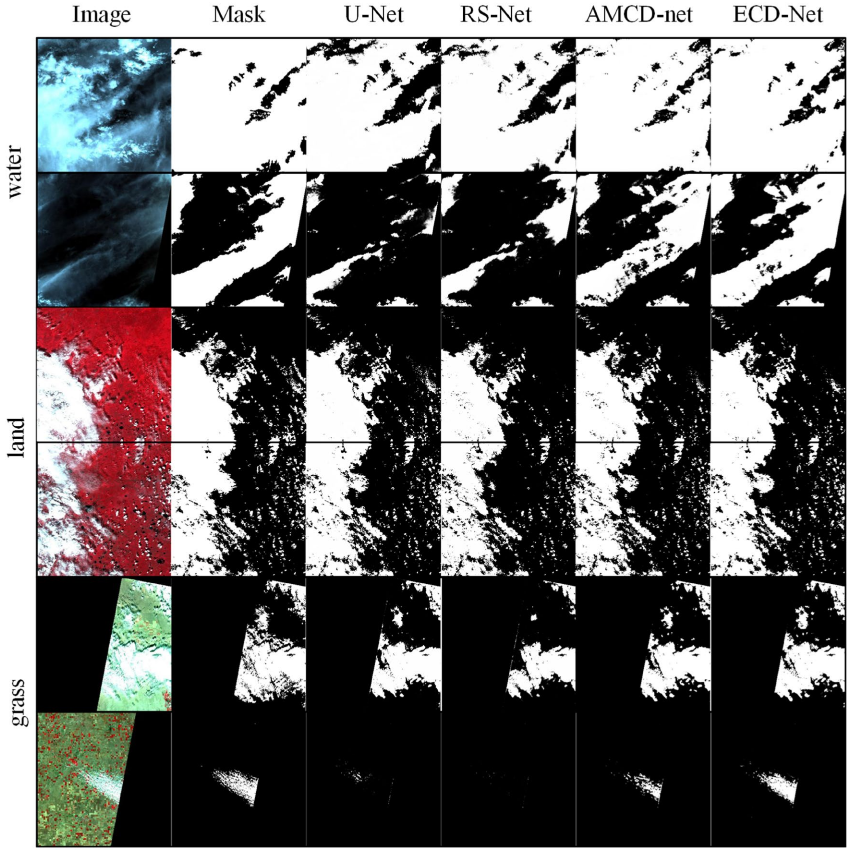

Figure 5. Visualization results on different methods.

In the land scene, where clouds and background are distinctly different in color and texture, all methods perform well in cloud detection. However, RS-Net's performance is somewhat lacking, with less precise segmentation of the clouds.

In the grass scene, U-Net and RS-Net show significant areas of missed detections, while AMCD-Net experiences only a few missed detections. ECD-Net delivers the best segmentation results in this scene, accurately identifying most of the cloud layers.

Overall, ECD-Net performs the best across all scenes, followed by AMCD-Net. U-Net and RS-Net show some limitations in both accuracy and detail handling.

4.3. Results from the Quantitative Evaluation on the SPARCS Dataset

As Table 2 shows that, consistent with its performance on the GF1-WHU dataset, ECD-Net achieves the best results in terms of accuracy (96.32%), Jaccard index (81.75%), and F1-Score (89.96%), fully demonstrating its exceptional performance in cloud detection tasks. However, despite having parameters and GFLOPS similar to other models, the inference speed of ECD-Net (33.70 FPS) is significantly lower.

Table 2. The results of different method.

Method	Parmas (M)	GFLOPs	FPS	Accuracy (%)	Jaccard (%)	F1-Score (%)
U-Net	17.263	89.955	61.43	96.11	80.93	89.46
RS-Net	9.389	38.183	63.61	96.07	80.56	89.23
AMCD-Net	10.025	37.187	49.45	96.14	81.22	89.64
ECD-Net (ours)	13.558	55.711	34.70	96.32	81.75	89.96

Overall, while ECD-Net excels in accuracy and detection capability, its slower inference speed may limit its applicability in real-time or high-efficiency scenarios. Therefore, future work should focus on optimizing the model for lightweight design and faster inference to enhance its practicality further.

4.4. Visualization Results of Different Methods on the SPARCS Dataset

As shown in Figure 6, all methods accurately detect the cloud regions in the first row, though ECD-Net and AMCD-Net exhibit a small number of false positives. In the second row, ECD-Net demonstrates the highest precision in capturing the boundaries and shapes of the cloud regions. In the third row, the edge detail detection of U-Net, RS-Net, and AMCD-Net appears somewhat coarse, while

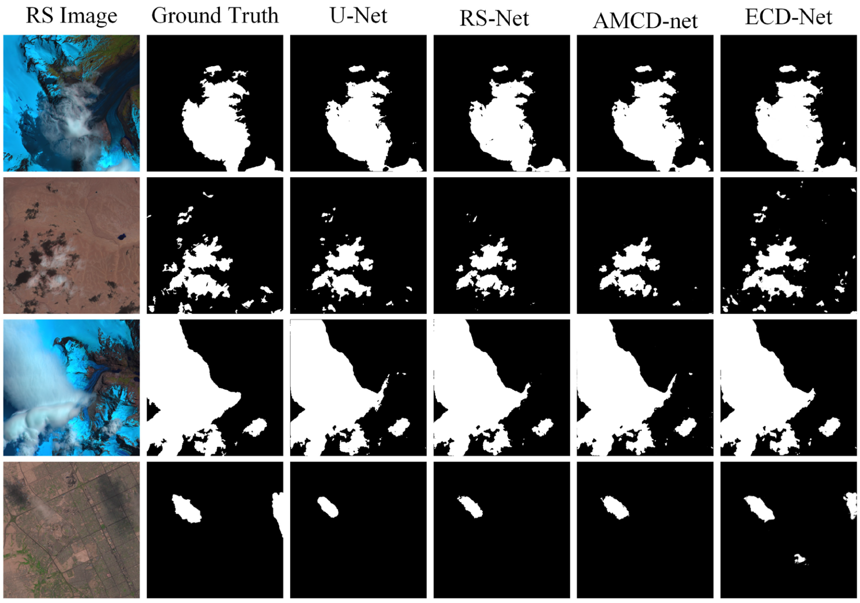

Figure 6. Visualization results on different methods.

ECD-Net more clearly restores the distribution details of the clouds. The fourth row illustrates a scene with sparse and isolated thin clouds, where ECD-Net accurately captures the thin cloud edges but shows false positives in the background of cloud-free areas. Overall, ECD-Net demonstrates superior performance in detecting edge details and thin clouds.

4.5. Ablation

To evaluate the effectiveness of multi-supervision in the ECD-Net model, we compared the performance of using BCE loss to supervise the cloud mask generated by the final layer with that of using multi-path supervision. The data in Table 3 reveals that integrating multi-path supervision yielded a 0.18% higher F1-score and 0.28% improvement in the Jaccard index compared to training with BCE loss alone. This improvement suggests that multi-path supervision provides more comprehensive guidance during training, significantly enhancing the model's performance in cloud detection tasks.

Table 3. Impact of loss function on the GF1-WHU dataset (%).

ECD-Net	$\mathcal{L}_{MS}$	Accuracy	Jaccard	F1-Score
✓	×	97.32	86.81	92.92
✓	✓	97.37	87.09	93.10

5. Conclusion

This study proposes an advanced cloud detection network, which is based on the U-Net architecture to address cloud identification in RS imagery. The method enhances the model's cloud detection capability in complex scenarios by introducing the MSDA and MHSA modules. Additionally, the designed multi-path supervision mechanism further improves the accuracy of cloud mask generation at multiple scales. Experimental findings using the GF1-WHU and SPARCS dataset demonstrate that the proposed model performs exceptionally well in complex scenarios, significantly improving cloud detection accuracy and showcasing strong potential for practical applications. Looking ahead, we aim to extend this method to other RS image datasets and explore its broad applications in cloud detection and cloud layer estimation. We also intend to investigate lightweight network architectures and apply knowledge distillation techniques to further reduce computational costs.

Acknowledgements

This research was supported by the Gansu Provincial Science and Technology Program (22YF7FA166).

Conflicts of Interest

The authors declare that there are no conflicts of interest associated with the publication of this paper.

References

[1] Zhang, Y., Rossow, W.B., Lacis, A.A., Oinas, V. and Mishchenko, M.I. (2004) Calculation of Radiative Fluxes from the Surface to Top of Atmosphere Based on ISCCP and Other Global Data Sets: Refinements of the Radiative Transfer Model and the Input Data. *Journal of Geophysical Research*: *Atmospheres*, **109**, D19105. https://doi.org/10.1029/2003jd004457

[2] Irish, R.R., Barker, J.L., Goward, S.N. and Arvidson, T. (2006) Characterization of the Landsat-7 ETM+ Automated Cloud-Cover Assessment (ACCA) Algorithm. *Photogrammetric Engineering & Remote Sensing*, **72**, 1179-1188. https://doi.org/10.14358/pers.72.10.1179

[3] Zhu, Z., Wang, S. and Woodcock, C.E. (2015) Improvement and Expansion of the Fmask Algorithm: Cloud, Cloud Shadow, and Snow Detection for Landsats 4-7, 8, and Sentinel 2 Images. *Remote Sensing of Environment*, **159**, 269-277. https://doi.org/10.1016/j.rse.2014.12.014

[4] Francis, A., Sidiropoulos, P. and Muller, J. (2019) CloudFCN: Accurate and Robust Cloud Detection for Satellite Imagery with Deep Learning. *Remote Sensing*, **11**, Article 2312. https://doi.org/10.3390/rs11192312

[5] Jeppesen, J.H., Jacobsen, R.H., Inceoglu, F. and Toftegaard, T.S. (2019) A Cloud Detection Algorithm for Satellite Imagery Based on Deep Learning. *Remote Sensing of Environment*, **229**, 247-259. https://doi.org/10.1016/j.rse.2019.03.039

[6] Li, W., Zou, Z. and Shi, Z. (2020) Deep Matting for Cloud Detection in Remote Sensing Images. *IEEE Transactions on Geoscience and Remote Sensing*, **58**, 8490-8502. https://doi.org/10.1109/tgrs.2020.2988265

[7] Wu, X., Shi, Z. and Zou, Z. (2021) A Geographic Information-Driven Method and a New Large Scale Dataset for Remote Sensing Cloud/Snow Detection. *ISPRS Journal of Photogrammetry and Remote Sensing*, **174**, 87-104. https://doi.org/10.1016/j.isprsjprs.2021.01.023

[8] Zhang, J., Wang, H., Wang, Y., Zhou, Q. and Li, Y. (2021) Deep Network Based on up and down Blocks Using Wavelet Transform and Successive Multi-Scale Spatial Attention for Cloud Detection. *Remote Sensing of Environment*, **261**, Article ID: 112483. https://doi.org/10.1016/j.rse.2021.112483

[9] Yang, J., Guo, J., Yue, H., Liu, Z., Hu, H. and Li, K. (2019) CDnet: CNN-Based Cloud Detection for Remote Sensing Imagery. *IEEE Transactions on Geoscience and Remote Sensing*, **57**, 6195-6211. https://doi.org/10.1109/tgrs.2019.2904868

[10] Guo, J., Yang, J., Yue, H., Tan, H., Hou, C. and Li, K. (2021) CDnetV2: CNN-Based Cloud Detection for Remote Sensing Imagery with Cloud-Snow Coexistence. *IEEE Transactions on Geoscience and Remote Sensing*, **59**, 700-713. https://doi.org/10.1109/tgrs.2020.2991398

[11] Li, Z., Shen, H., Cheng, Q., Liu, Y., You, S. and He, Z. (2019) Deep Learning Based Cloud Detection for Medium and High Resolution Remote Sensing Images of Different Sensors. *ISPRS Journal of Photogrammetry and Remote Sensing*, **150**, 197-212. https://doi.org/10.1016/j.isprsjprs.2019.02.017

[12] Li, X., Yang, X., Li, X., Lu, S., Ye, Y. and Ban, Y. (2022) GCDB-UNet: A Novel Robust Cloud Detection Approach for Remote Sensing Images. *Knowledge-Based Systems*, **238**, Article ID: 107890. https://doi.org/10.1016/j.knosys.2021.107890

[13] Wu, K., Xu, Z., Lyu, X. and Ren, P. (2022) Cloud Detection with Boundary Nets. *ISPRS Journal of Photogrammetry and Remote Sensing*, **186**, 218-231.

https://doi.org/10.1016/j.isprsjprs.2022.02.010

[14] Zhai, H. and Xue, L. (2024) AMCD-Net: An Effective Attention-Aided Multi-level Cloud Detection Network for Optical Satellite Imagery. *IEEE Transactions on Geoscience and Remote Sensing*, **62**, 1-21. https://doi.org/10.1109/tgrs.2024.3372589

[15] Martins, B.J., Arrais, J.M., Cerentini, A., Wangenheim, A.v., Neto, G.P.R. and Mantelli, S. (2023) Segmentation and Classification of Individual Clouds in Images Captured with Horizon-Aimed Cameras for Nowcasting of Solar Irradiance Absorption. *American Journal of Climate Change*, **12**, 628-654. https://doi.org/10.4236/ajcc.2023.124027

[16] Matsunobu, L.M., Pedro, H.T.C. and Coimbra, C.F.M. (2021) Cloud Detection Using Convolutional Neural Networks on Remote Sensing Images. *Solar Energy*, **230**, 1020-1032. https://doi.org/10.1016/j.solener.2021.10.065

[17] Cao, H., Wang, Y., Chen, J., Jiang, D., Zhang, X., Tian, Q., et al. (2023) Swin-Unet: Unet-Like Pure Transformer for Medical Image Segmentation. In: Karlinsky, L., Michaeli, T. and Nishino, K., Eds., *Computer Vision—ECCV* 2022 *Workshops*, Springer Nature Switzerland, 205-218. https://doi.org/10.1007/978-3-031-25066-8_9

[18] Xie, E., Wang, W., Yu, Z., Anandkumar, A., Alvarez, J.M. and Luo, P. (2021) SegFormer: Simple and Efficient Design for Semantic Segmentation with Transformers. *Proceedings of the* 35*th Conference on Neural Information Processing Systems* (*NeurIPS*'21), Vancouver, 6-14 December 2021, 12077-12090.

[19] Liu, X., Peng, H., Zheng, N., Yang, Y., Hu, H. and Yuan, Y. (2023) EfficientViT: Memory Efficient Vision Transformer with Cascaded Group Attention. 2023 *IEEE*/*CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), Vancouver, 17-24 June 2023, 14420-14430. https://doi.org/10.1109/cvpr52729.2023.01386

[20] Yeom, S.K., and von Klitzing, J. (2023) U-MixFormer: UNet-Like Transformer with Mix-Attention for Efficient Semantic Segmentation. arXiv: 2312.06272.

[21] Ma, J., Li, F., and Wang, B. (2024) U-Mamba: Enhancing Long-Range Dependency for Biomedical Image Segmentation. arXiv: 2401.04722.

[22] Liu, M., Dan, J., Lu, Z., Yu, Y., Li, Y., and Li, X. (2024) CM-UNet: Hybrid CNN-Mamba UNet for Remote Sensing Image Semantic Segmentation. arXiv: 2405.10530.

[23] Jiao, J., Tang, Y., Lin, K., Gao, Y., Ma, A.J., Wang, Y., et al. (2023) Dilateformer: Multi-Scale Dilated Transformer for Visual Recognition. *IEEE Transactions on Multimedia*, **25**, 8906-8919. https://doi.org/10.1109/tmm.2023.3243616

[24] Vaswani, A., Shazeer, N.M., Parmar, N., Uszkoreit, J., Jones, L., et al. (2017) Attention Is All You Need. *Proceedings of the* 31*st Neural Information Processing Systems* (*NeurIPS*), Long Beach, 4-9 December 2017, 5998-6008.

[25] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., et al. (2021) An Image is Worth 16 × 16 Words: Transformers for Image Recognition at Scale. *Proceedings of the International Conference on Learning Representations* (*ICLR*).

[26] Chu, X., Tian, Z., Zhang, B., Wang, X. and Shen, C. (2021) Conditional Positional Encodings for Vision Transformers. *Proceedings of the International Conference on Learning Representations* (*ICLR*).

[27] de Boer, P., Kroese, D.P., Mannor, S. and Rubinstein, R.Y. (2005) A Tutorial on the Cross-Entropy Method. *Annals of Operations Research*, **134**, 19-67. https://doi.org/10.1007/s10479-005-5724-z

[28] Li, Z., Shen, H., Li, H., Xia, G., Gamba, P. and Zhang, L. (2017) Multi-Feature Combined Cloud and Cloud Shadow Detection in Gaofen-1 Wide Field of View Imagery. *Remote Sensing of Environment*, **191**, 342-358.

https://doi.org/10.1016/j.rse.2017.01.026

[29] Hughes, M. and Hayes, D. (2014) Automated Detection of Cloud and Cloud Shadow in Single-Date Landsat Imagery Using Neural Networks and Spatial Post-Processing. *Remote Sensing*, **6**, 4907-4926. https://doi.org/10.3390/rs6064907

[30] Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., et al. (2019) PyTorch: An Imperative Style, High-Performance Deep Learning Library. *Proceedings of the* 33*rd International Conference on Neural Information Processing Systems*, Vancouver, 8-14 December 2019, 8026-8037.

[31] Loshchilov, I., and Hutter, F. (2019) Decoupled Weight Decay Regularization. *Proceedings of the International Conference on Learning Representations* (*ICLR*).

[32] Long, J., Shelhamer, E. and Darrell, T. (2015) Fully Convolutional Networks for Semantic Segmentation. 2015 *IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR*), Boston, 7-12 June 2015, 3431-3440. https://doi.org/10.1109/cvpr.2015.7298965

[33] Powers, D.M.W. (2020) Evaluation: From Precision, Recall and F-Measure to ROC, Informedness, Markedness and Correlation. arXiv: 2010.16061.