SEGREGATION AND CONTEXT AGGREGATION NETWORK FOR REAL-TIME CLOUD SEGMENTATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Cloud segmentation from intensity images is a pivotal task in atmospheric science and computer vision, aiding weather forecasting and climate analysis. Ground-based sky/cloud segmentation extracts clouds from images for further feature analysis. Existing methods struggle to balance segmentation accuracy and computational efficiency, limiting real-world deployment on edge devices, so we introduce SCANet, a novel lightweight cloud segmentation model featuring Segregation and Context Aggregation Module (SCAM), which refines rough segmentation maps into weighted sky and cloud features processed separately. SCANet achieves state-of-the-art performance while drastically reducing computational complexity. SCANet-large (4.29M) achieves comparable accuracy to state-of-the-art methods with 70.9% fewer parameters. Meanwhile, SCANet-lite (90K) delivers 1390 fps in FP16, surpassing real-time standards. Additionally, we propose an efficient pre-training strategy that enhances performance even without ImageNet pre-training.

Keywords: cloud segmentation, machine learning, segregation and context aggregation module.

1 INTRODUCTION

Understanding cloud-sky relationships is crucial for climate modeling, solar energy forecasting, and extreme weather prediction. Advances in computer vision and machine learning have improved meteorology estimation (1; 2; 3) and weather prediction (4; 5; 6), offering insights into cloud status. While satellites provide valuable cloud data, they are costly and storage-intensive. Ground-based sky/cloud segmentation (7; 8; 9), supported by datasets like SWIMSEG (10), SWINSEG (11), and SWINySEG (12), offers a cost-effective, high-resolution alternative, enhancing climate applications.

Sky/cloud segmentation, a binary semantic task, has evolved with fully convolutional networks (FCN) (13). Real-time segmentation strategies include (a) lightweight backbones and decoders, e.g., DeepLab (14; 15), and (b) encoder-decoder architectures like ICNet (16) and BiseNet (17; 18). However, existing methods either lose information due to attention mechanisms or suffer from slow inference, limiting real-time applications. A detailed discussion of related works, including their limitations, can be found in Appendix A.1, further motivating the design of SCANet.

To address these challenges, we propose SCANet, a lightweight yet effective model integrating MobileNetV2 (19) and EfficientNet-B0 (20) with a Segregation and Context Aggregation module (SCAM). SCANet, following a U-Net (21) structure, refines cloud-sky features via SCAM decoders. Unlike prior methods, SCAM enhances feature separation and aggregation, while supervision at the last three stages accelerates convergence.

By improving segmentation accuracy and efficiency, SCANet supports climate modeling, renewable energy forecasting, and extreme weather monitoring, demonstrating a direct pathway from machine learning to climate impact. Further details on SCANet's contributions to climate change mitigation and adaptation are discussed in Appendix A.8. The main contributions of our SCANet are twofold:

- SCANet, a lightweight CNN-based model, achieves state-of-the-art performance with 70.68% fewer parameters while exceeding real-time standards, incorporating a new pre-training strategy for sky/cloud segmentation when ImageNet pre-training is unavailable.
- A novel SCAM decoder with segregated branches for information processing, enabling precise segmentation while maintaining real-time efficiency.

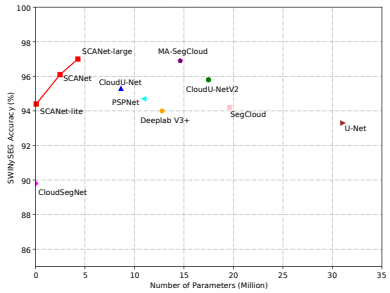

Figure 1: #Params vs. SWINySEG Accuracy. Our proposed SCANet model successfully achieves a balance between the model size and accuracy. SCANet-large can achieve 97.0% of accuracy in SWINySEG with 4.29 million parameters, while SCANet-lite can achieve 94.4% of accuracy with only 90k parameters.

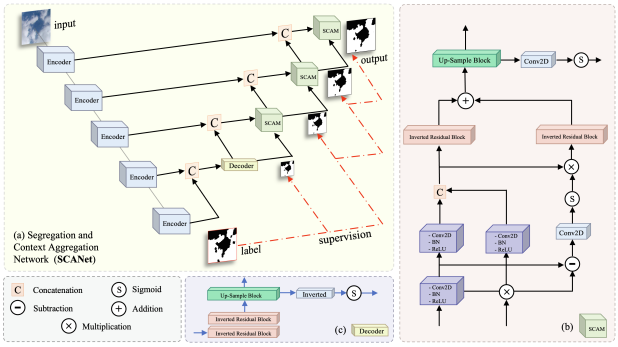

Figure 2: The overall architecture of SCANet and SCAM. (a) presents the pipeline of SCANet; (b) details the design of SCAM; (c) depicts the decoder structure preceding the SCAM modules. Additionally, The architecture of inverted residual block (19) is demonstrated in Fig. 4 in Appendix A.6. The Up-sample block consists of an inverted residual block paired with a bilinear Up-sample layer.

2 SCANET

Our SCANet is designed based on the U-Net (21) structure. We use different backbone networks to extract high-dimensional features in our experiments. The architecture of SCANet is shown in Fig. 2 (a). We equip our SCANet with MobileNetV2 (19) and EfficientNet-B0 (20) for different settings. We also propose SWINySEG pre-training (SWPT) for SCANet-lite, demonstrated in Appendix A.3.

2.1 SEGREGATION AND CONTEXT AGGREGATION MODULE (SCAM)

The Segregation and Context Aggregation Module (SCAM) is a lightweight decoder for sky/cloud segmentation, as shown in Fig. 2 (b), operating in the 2nd, 3rd, and 4th stages. Given the binary nature of sky/cloud segmentation, SCAM processes these categories separately by first segregating features based on rough segmentation results from the previous stage and then aggregating them. It takes two inputs: a concatenation of the U-Net shortcut and feature maps from the previous layer, c_{i-1}, and the segmentation prediction from the prior stage, s_{i-1}. The main branch is formulated as $f_i = \mathrm{Cat}(\mathrm{Conv}(c_{i-1}), \mathrm{Conv}(c_{i-1} \times s_{i-1}))$, where f_i represents foreground (sky) features, and Cat denotes channel-wise concatenation. The background mask is computed as $m_i = \mathrm{Sigmoid}(\mathrm{Conv}(c_{i-1} \times (1 - s_{i-1})))$. To extract background features, we apply the background mask to the core branch, formulated as $b_i = f_i \times m_i$, improving background representation. Finally, element-wise addition aggregates f_i and b_i, producing the output o_i as $o_i = \mathrm{Conv}(\mathrm{UpSample}(\mathrm{Conv}(b_i) + \mathrm{Conv}(f_i)))$, while the stage prediction is obtained through a convolutional layer followed by a sigmoid activation: $s_i = \mathrm{Sigmoid}(\mathrm{Conv}(o_i))$.

2.2 LOSS FUNCTIONS

In our research, We use binary cross entropy (BCE) and Intersection of Union (IOU) loss as the loss function in the training of SCANet. These loss functions can be defined as follows:

$$\mathcal{L}_{\mathrm{bce}}(p, y) = -\frac{1}{N} * \sum_{j=1}^{N} (y_j * \log p_j + (1 - y_j) * \log (1 - p_j)) \tag{1}$$

$$\mathcal{L}_{\mathrm{iou}}(p, y) = 1 - \frac{1}{n} \sum_{j=1}^{N} \left(\frac{y_j \times p_j}{y_j + p_j - y_j \times p_j}\right) \tag{2}$$

then our total training loss under deep supervision can be formulated as:

$$\mathcal{L}(p, y) = \sum_{i=1}^{4} \alpha_i * (\mathcal{L}_{\mathrm{bce}}(p_i, y_i) + \mathcal{L}_{\mathrm{iou}}(p_i, y_i)) \tag{3}$$

in which α_i represents the coefficient of i th SCAM or decoder.

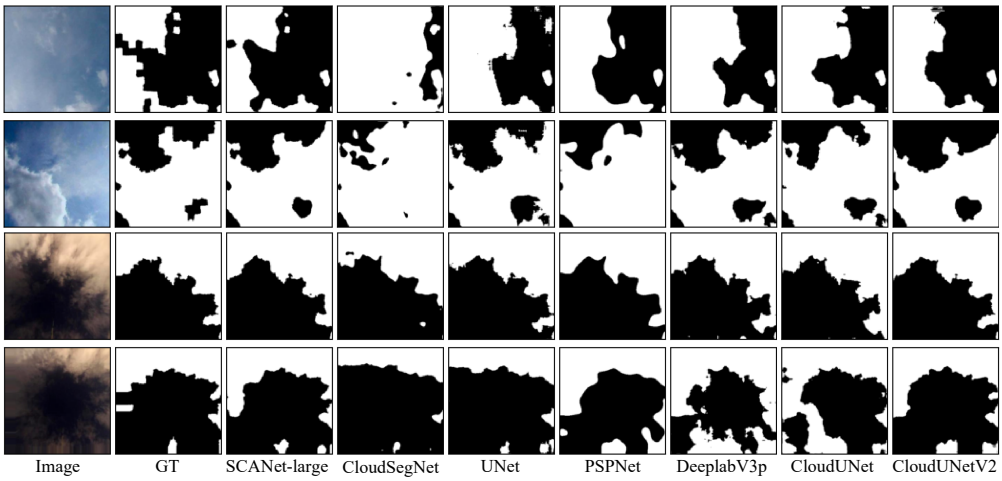

Figure 3: Qualitative comparison of SCANet-large with state-of-the-art approaches on day-time (rows 1–2) and night-time (rows 3–4) images from the SWINySEG dataset.

3 Experiments & Results

We conduct experiments on Singapore Whole Sky Nychthemeron Image SEGmentation Database (SWINySEG), see Appendix A.4 for details. Our experiment setting is described in Appendix A.5.

3.1 Metrics

In our experiments, we evaluate SCANet using six widely used metrics: accuracy, precision, recall, F-score, error rate, and MIoU. The F-score, which reflects overall model performance, is the harmonic mean of precision and recall, given by $\frac{2\times\text{Precision}\times\text{Recall}}{\text{Precision}+\text{Recall}}$. Precision is defined as $\frac{\text{TP}}{\text{TP+FP}}$, recall as $\frac{\text{TP}}{\text{TP+FN}}$, and error rate as $\frac{\text{FP+FN}}{\text{P+N}}$. Besides, MIoU, a common metric is calculated as $\text{MIoU} = \frac{\text{miou}_+ + \text{miou}_-}{2}$, where $\text{miou}_+ = \frac{\text{TP}}{\text{FN+FP+TP}}$ and $\text{miou}_- = \frac{\text{TN}}{\text{TN+FN+FP}}$, respectively.

3.2 Qualitative Evaluation

The qualitative comparison of SCANet-large with six prior methods (12; 22; 23; 21; 24; 25) is shown in Fig. 3. The leftmost columns present source images and ground truths, with day-time samples in the first two rows and night-time images in the last two. Cloud-sky boundaries in the first row challenge prior methods—CloudSegNet misclassifies sky as cloud, while others miss small patches. SCANet-large, however, accurately segments both. In the second row, it correctly classifies three small sky patches, unlike previous models that confuse sky and cloud. Night-time segmentation is even harder due to limited training data. In the first night-time row, complex cloud structures degrade prior methods' performance, but SCANet-large remains accurate, highlighting its advantages in atmospheric science applications. Additional comparisons are provided in Fig. 6 in Appendix A.6.

3.3 Quantitative Evaluation

Table 1 presents the quantitative evaluation of SCANet on day-time, night-time, and day+night time SWINySEG datasets, comparing it with state-of-the-art methods. We reference Zhang et al. (29) for prior results and ensure a fair comparison by maintaining the same settings. SCANet-large achieves the highest accuracy (0.970) and precision (0.971), outperforming 8 prior methods, including MA-SegCloud (0.969 accuracy, 0.970 precision), despite having only 4.29 million parameters—a 70.68% reduction compared to MA-SegCloud (14.63 million). The standard SCANet achieves competitive accuracy (0.960) and MIoU (0.900), while SCANet-lite, with just 0.09 million parameters, attains 0.945 accuracy, surpassing larger models like SegCloud (19.61 million parameters, 0.942 accuracy). These results demonstrate SCANet's efficiency and performance balance, with SCANet-large achieving state-of-the-art accuracy and precision using much fewer parameters. Precision-Recall (PR) curves in Fig. 7 in Appendix A.6 illustrate all methods' overall performance.

Table 1: Comparison with other state-of-the-art methods on day time, night time, and day+night time images. We highlight the optimal and suboptimal methods in **bold** and underline, respectively. "-" indicates that the corresponding metric was not reported in the original paper.

Methods	#Params	Day-time					Night-time					Day+Night time				
		Acc.	Prec.	Rec.	F1	MIoU	Acc.	Prec.	Rec.	F1	MIoU	Acc.	Prec.	Rec.	F1	MIoU
General Semantic Segmentation Models																
U-Net (21)	31.05M	0.933	0.938	0.919	0.928	0.844	0.933	0.938	0.919	0.928	0.844	0.933	0.938	0.919	0.928	0.844
PSPNet (24)	20.95M	0.947	0.949	0.935	0.942	0.873	0.947	0.950	0.935	0.942	0.873	0.947	0.950	0.935	0.942	0.873
DeeplabV3+ (25)	12.80M	0.940	0.941	0.920	0.936	0.860	0.940	0.941	0.932	0.936	0.860	0.940	0.941	0.932	0.936	0.860
Special Designed Sky/Cloud Segmentation Models																
CloudSegNet (12)	0.005M	0.898	0.920	0.876	0.898	0.777	0.898	0.920	0.876	0.898	0.777	0.898	0.920	0.876	0.898	0.777
SegCloud (26)	19.61M	0.941	0.953	0.934	0.943	0.889	0.955	0.936	0.960	0.948	0.912	0.942	0.952	0.936	0.944	0.891
UCloudNet (27)	-	0.940	0.920	0.940	0.930	-	0.960	0.950	0.950	0.950	-	0.940	0.920	0.940	0.930	-
DDUNet (28)	0.33M	0.953	0.953	-	-	0.882	0.954	0.951	-	-	0.900	0.953	0.952	-	-	0.884
CloudU-Net (22)	8.64M	0.953	0.949	0.955	0.948	0.885	0.953	0.949	0.947	0.948	0.885	0.953	0.949	0.947	0.948	0.885
CloudU-NetV2 (23)	17.48M	0.958	0.955	0.952	0.953	0.895	0.958	0.955	0.952	0.953	0.895	0.958	0.955	0.952	0.953	0.900
MA-SegCloud (29)	14.63M	0.969	0.971	**0.970**	**0.970**	**0.940**	0.969	0.960	**0.970**	0.965	**0.940**	0.969	0.970	**0.970**	**0.970**	**0.940**
SCANet-lite	0.09M	0.944	0.936	0.944	0.940	0.865	0.944	0.936	0.944	0.940	0.865	0.944	0.936	0.944	0.940	0.865
SCANet	2.49M	0.961	0.955	0.958	0.957	0.901	0.961	0.955	0.958	0.957	0.902	0.961	0.955	0.958	0.957	0.902
SCANet-large	4.29M	**0.970**	**0.971**	0.960	0.966	0.923	**0.970**	**0.971**	0.960	**0.966**	0.923	**0.970**	**0.971**	0.960	0.966	0.923

Methods	FP32		FP16	
	FPS	Latency	FPS	Latency
SCANet-lite	750	1.3 ms	1390	0.7 ms
SCANet	465	2.1 ms	1124	0.8 ms
SCANet-large	299	3.3 ms	392	2.6 ms

Table 2: Inference latency and FPS of SCANet configurations on an NVIDIA Tesla V100-SXM2 16GB GPU. Models were deployed using TensorRT with FP32 and FP16 precision. Inference latency is measured as the average processing time per image over 1000 inferences.

3.4 Ablation Study

To assess our proposed modules, backbone networks, loss functions, and pre-training strategies, we conduct an ablation study shown in Table 3. The baseline U-Net with inverted residual blocks has 0.32M parameters, achieving 92.7% accuracy and 83.2% MIoU with BCE loss. Replacing its backbone with MobileNetV2-lite and adding SCAM (without the Right Branch) boosts accuracy to 93.6% (No. 2) with only 0.09M parameters. The complete SCAM (No. 3) further improves performance, while SWPT provides minor gains. BCE+IOU loss surpasses IOU-only in accuracy and MIoU. Finally, we evaluate MobileNetV2 and EfficientNet-B0 with BCE+IOU loss. To complement quantitative results, Fig. 9 in Appendix A.6 provides visualizations of eight key experiments (No. 1, 2, 3, 4, 5, 6, 8, 10), alongside PR and F-Measure curves in Fig. 8 within Appendix A.6.

Table 3: Ablation study on different module compositions, loss functions, backbone networks, and pre-training strategies. SWPT indicates SWINySEG-based pre-training and INPT is ImageNet-based pre-training. We build a light-weight U-Net with 0.32M parameters as the baseline model.

No.	Backbone	SCAM Configs		Loss Functions		Pre-training		#Params	SWINySEG				
		L Branch	R Branch	BCE	IOU	SWPT	INPT		Accuracy	Precision	Recall	F-score	MIoU
1	baseline	✗	✗	✓	✗	✗	✗	0.32M	0.927	0.934	0.910	0.922	0.832
2	MobileNetV2-lite	✓	✗	✓	✗	✗	✗	**0.09M**	0.936	0.940	0.925	0.932	0.850
3	MobileNetV2-lite	✓	✓	✓	✗	✗	✗	**0.09M**	0.943	0.944	0.932	0.938	0.863
4	MobileNetV2-lite	✓	✓	✓	✗	✓	✗	**0.09M**	0.944	0.943	0.933	0.938	0.863
5	MobileNetV2-lite	✓	✓	✗	✓	✓	✗	**0.09M**	0.940	0.941	0.934	0.938	0.856
6	MobileNetV2-lite	✓	✓	✓	✓	✓	✗	**0.09M**	0.944	0.936	0.944	0.940	0.865
7	MobileNetV2	✓	✓	✓	✗	✗	✓	2.49M	0.960	0.958	0.951	0.955	0.898
8	MobileNetV2	✓	✓	✓	✓	✗	✓	2.49M	0.961	0.955	0.958	0.957	0.902
9	EfficientNet-B0	✓	✓	✓	✗	✗	✓	4.29M	0.969	**0.972**	0.956	0.964	0.919
10	EfficientNet-B0	✓	✓	✓	✓	✗	✓	4.29M	**0.970**	0.971	**0.960**	**0.966**	**0.923**

4 Conclusion

In this paper, we introduce SCANet, a real-time lightweight cloud segmentation model that reduces parameters by 70.9% while maintaining state-of-the-art performance. The SCANet-large configuration achieves 392 fps in FP16 after TensorRT deployment, whereas SCANet-lite with only 0.09M reaches 1390 fps. We also propose an efficient pre-training strategy that enhances segmentation accuracy when ImageNet pre-training is unavailable. Extensive evaluations with prior advanced methods, confirm SCANet's superior accuracy and inference speed, exceeding real-time standards.

REFERENCES

[1] S. Dev, F. M. Savoy, Y. H. Lee, S. Winkler, Estimating Solar Irradiance Using Sky Imagers, Atmospheric Measurement Techniques 12 (10) (2019) 5417–5429.

[2] D. Tulpan, C. Bouchard, K. Ellis, C. Minwalla, Detection of clouds in sky/cloud and aerial images using moment based texture segmentation, in: Proceedings of the International Conference on Unmanned Aircraft Systems (ICUAS), 2017, pp. 1124–1133.

[3] N. Akrami, K. Ziarati, S. Dev, Graph-based Local Climate Classification in Iran, International Journal of Climatology 42 (3) (2022) 1337–1353.

[4] S. Manandhar, S. Dev, Y. H. Lee, Y. S. Meng, S. Winkler, A Data-Driven Approach for Accurate Rainfall Prediction, IEEE Transactions on Geoscience and Remote Sensing 57 (11) (2019) 9323–9331.

[5] H. Wang, M. S. Pathan, Y. H. Lee, S. Dev, Day-ahead Forecasts of Air Temperature, in: 2021 IEEE USNC-URSI Radio Science Meeting (Joint with AP-S Symposium), IEEE, 2021, pp. 94–95.

[6] B. McNicholl, Y. H. Lee, A. G. Campbell, S. Dev, Evaluating the Reliability of Air Temperature From ERA5 Reanalysis Data, IEEE Geoscience and Remote Sensing Letters 19 (2021) 1–5.

[7] M. Jain, I. Gollini, M. Bertolotto, G. McArdle, S. Dev, An Extremely-Low Cost Ground-Based Whole Sky Imager, in: Proceedings of the IEEE International Geoscience and Remote Sensing Symposium (IGARSS), IEEE, 2021, pp. 8209–8212.

[8] S. Dev, F. M. Savoy, Y. H. Lee, S. Winkler, Design of low-cost, compact and weather-proof whole sky imagers for high-dynamic-range captures, in: Proceedings of the IEEE International Geoscience and Remote Sensing Symposium (IGARSS), IEEE, 2015, pp. 5359–5362.

[9] S. Dev, F. M. Savoy, Y. H. Lee, S. Winkler, High-Dynamic-Range Imaging for Cloud Segmentation, Atmospheric Measurement Techniques 11 (4) (2018) 2041–2049.

[10] S. Dev, Y. H. Lee, S. Winkler, Color-Based Segmentation of Sky/Cloud Images From Ground-Based Cameras, IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing 10 (1) (2016) 231–242.

[11] S. Dev, F. M. Savoy, Y. H. Lee, S. Winkler, Nighttime sky/cloud image segmentation, in: Proceedings of the IEEE International Conference on Image Processing (ICIP), IEEE, 2017, pp. 345–349.

[12] S. Dev, A. Nautiyal, Y. H. Lee, S. Winkler, CloudSegNet: A Deep Network for Nychthemeron Cloud Image Segmentation, IEEE Geoscience and Remote Sensing Letters (GRSL) 16 (12) (2019) 1814–1818.

[13] J. Long, E. Shelhamer, T. Darrell, Fully Convolutional Networks for Semantic Segmentation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 3431–3440.

[14] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, A. L. Yuille, Semantic Image Segmentation with Deep Convolutional Nets and Fully Connected CRFs, arXiv preprint arXiv:1412.7062 (2014).

[15] L. C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, A. L. Yuille, DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs, IEEE Transactions on Pattern Analysis and Machine Intelligence (IEEE T-PAMI) 40 (4) (2018) 834–848.

[16] H. Zhao, X. Qi, X. Shen, J. Shi, J. Jia, ICNet for Real-Time Semantic Segmentation on High-Resolution Images, in: Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 405–420.

[17] C. Yu, J. Wang, C. Peng, C. Gao, G. Yu, N. Sang, BiSeNet: Bilateral Segmentation Network for Real-time Semantic Segmentation, in: Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 325–341.

[18] C. Yu, C. Gao, J. Wang, G. Yu, C. Shen, N. Sang, BiSeNet V2: Bilateral Network with Guided Aggregation for Real-Time Semantic Segmentation, International Journal of Computer Vision (IJCV) 129 (11) (2021) 3051–3068.

[19] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, L.-C. Chen, MobileNetV2: Inverted Residuals and Linear Bottlenecks , in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 4510–4520.

[20] M. Tan, Q. Le, EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks, in: Proceedings of the International Conference on Machine Learning (PMLR), PMLR, 2019, pp. 6105–6114.

[21] O. Ronneberger, P. Fischer, T. Brox, U-Net: Convolutional Networks for Biomedical Image Segmentation, in: Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, 2015, pp. 234–241.

[22] C. Shi, Y. Zhou, B. Qiu, D. Guo, M. Li, CloudU-Net: A Deep Convolutional Neural Network Architecture for Daytime and Nighttime Cloud Images' Segmentation, IEEE Geoscience and Remote Sensing Letters (GRSL) 18 (10) (2020) 1688–1692.

[23] C. Shi, Y. Zhou, B. Qiu, CloudU-Netv2: A Cloud Segmentation Method for Ground-Based Cloud Images Based on Deep Learning, Neural Processing Letters 53 (4) (2021) 2715–2728.

[24] H. Zhao, J. Shi, X. Qi, X. Wang, J. Jia, Pyramid Scene Parsing Network, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017.

[25] L.-C. Chen, G. Papandreou, F. Schroff, H. Adam, Rethinking Atrous Convolution for Semantic Image Segmentation, arXiv preprint arXiv:1706.05587 (2017).

[26] W. Xie, D. Liu, M. Yang, S. Chen, B. Wang, Z. Wang, Y. Xia, Y. Liu, Y. Wang, C. Zhang, SegCloud: A novel cloud image segmentation model using a deep convolutional neural network for ground-based all-sky-view camera observation, Atmospheric Measurement Techniques 13 (4) (2020) 1953–1961.

[27] Y. Li, H. Wang, S. Wang, Y. H. Lee, M. S. Pathan, S. Dev, UCloudNet: A Residual U-Net with Deep Supervision for Cloud Image Segmentation, in: IGARSS 2024-2024 IEEE International Geoscience and Remote Sensing Symposium, IEEE, 2024, pp. 5553–5557.

[28] Y. Li, H. Wang, J. Xu, P. Wu, Y. Xiao, S. Wang, S. Dev, DDUNet: Dual Dynamic U-Net for Highly-Efficient Cloud Segmentation, arXiv preprint arXiv:2501.15385 (2025).

[29] L. Zhang, W. Wei, B. Qiu, A. Luo, M. Zhang, X. Li, A Novel Ground-Based Cloud Image Segmentation Method Based on a Multibranch Asymmetric Convolution Module and Attention Mechanism, Remote Sensing 14 (16) (2022) 3970.

[30] C. N. Long, J. M. Sabburg, J. Calbó, D. Pagès, Retrieving Cloud Characteristics from Ground-Based Daytime Color All-Sky Images, Journal of Atmospheric and Oceanic Technology 23 (5) (2006) 633–652.

[31] S. Dev, Y. H. Lee, S. Winkler, Systematic study of color spaces and components for the segmentation of sky/cloud images, in: Proceedings of the IEEE International Conference on Image Processing (ICIP), IEEE, 2014, pp. 5102–5106.

[32] J. Yang, W. Lv, Y. Ma, W. Yao, Q. Li, An Automatic Groundbased Cloud Detection Method based on Local Threshold Interpolation, Acta Meteorologica Sinica 68 (6) (2010) 1007–1017.

[33] S. Dev, S. Manandhar, Y. H. Lee, S. Winkler, Multi-label Cloud Segmentation Using a Deep Network, in: Proceedings of the USNC-URSI Radio Science Meeting (Joint with AP-S Symposium), IEEE, 2019, pp. 113–114.

[34] X. Wang, R. Girshick, A. Gupta, K. He, Non-local neural networks, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 7794–7803.

[35] S. Woo, J. Park, J.-Y. Lee, I. S. Kweon, CBAM: Convolutional Block Attention Module, in: Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 3–19.

[36] J. Hu, L. Shen, G. Sun, Squeeze-and-Excitation Networks, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 7132–7141.

[37] S. Liu, J. Zhang, Z. Zhang, X. Cao, T. S. Durrani, TransCloudSeg: Ground-based cloud image segmentation with transformer, IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing 15 (2022) 6121–6132.

[38] A. L. De Souza, P. Shokri, Residual u-net with attention for detecting clouds in satellite imagery (2023).

[39] Y. Guo, X. Cao, B. Liu, M. Gao, Cloud detection for satellite imagery using attention-based u-net convolutional neural network, Symmetry 12 (6) (2020) 1056.

[40] P. K. Buttar, M. K. Sachan, Semantic segmentation of clouds in satellite images based on u-net++ architecture and attention mechanism, Expert Systems with Applications 209 (2022) 118380.

[41] Z. Zhou, M. M. R. Siddiquee, N. Tajbakhsh, J. Liang, Unet++: Redesigning skip connections to exploit multiscale features in image segmentation, IEEE transactions on medical imaging 39 (6) (2019) 1856–1867.

[42] M. Partio, L. Hieta, A. Kokkonen, CloudCast–Total Cloud Cover Nowcasting with Machine Learning, arXiv preprint arXiv:2410.21329 (2024).

[43] Y. Li, H. Wang, Z. Li, S. Wang, S. Dev, G. Zuo, DAANet: Dual Attention Aggregating Network for Salient Object Detection, in: IEEE International Conference on Robotics and Biomimetics (ROBIO), IEEE, 2023, pp. 1–7.

[44] Y. Li, H. Wang, A. Katsaggelos, CPDR: Towards Highly-Efficient Salient Object Detection via Crossed Post-decoder Refinement, in: 35th British Machine Vision Conference 2024, BMVC 2024, Glasgow, UK, November 25-28, 2024, BMVA, 2024.

[45] K. He, X. Zhang, S. Ren, J. Sun, Deep Residual Learning for Image Recognition, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 770–778.

[46] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, L. Fei-Fei, Imagenet: A large-scale hierarchical image database, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, 2009, pp. 248–255.

A APPENDIX

A.1 RELATED WORK

Earlier methods in cloud segmentation relied heavily on traditional techniques, utilizing color features, pre-defined convolution filters, and edge detection operators (30; 31; 32). These methods, however, often struggled with capturing fine-grained details, resulting in sub-par segmentation accuracy. Additionally, their inability to model probabilistic relationships limited their effectiveness when handling unseen data.

Since deep learning emerged, CNN-based approaches have largely dominated the field of sky/cloud segmentation research. For instance, Dev *et al.* (12) developed CloudSegNet, a fully convolutional network that applies down-sampling and up-sampling processes to extract high-dimensional feature maps, resulting in segmentation masks with better boundary accuracy. Dev *et al.* (33) introduced a multi-label segmentation method, classifying cloud images into thin clouds, thick clouds, and sky categories, using a U-Net architecture to enhance segmentation precision.

In 2021, Shi *et al.* (22) presented CloudU-Net, a U-Net-based model that integrates fully connected conditional random field (CRF) layers for refined post-processing. This model was further improved into CloudU-NetV2 (23), which included non-local attention (34) to better capture long-range dependencies, thus improving segmentation accuracy. Nonetheless, adding multiple attention modules significantly increased computational costs due to additional matrix multiplications. Subsequently, Zhang *et al.* (29) introduced MA-SegCloud, which incorporates the convolutional block attention module (CBAM) (35), squeeze-and-excitation module (SEM) (36), and asymmetric convolution mechanisms to boost segmentation performance.

Recently, Transformer-based models have been explored to handle long-range dependencies in sky/cloud segmentation. Liu *et al.* (37) proposed TransCloudSeg, a hybrid model that merges CNN-based encoders with Transformer-based feature extractors, utilizing a Heterogeneous Fusion Module (HFM) to combine outputs from CNN and Transformer branches, thereby demonstrating superior segmentation results. Additionally, research by Souza *et al.* (38) and Guo *et al.* (39) underscores the benefits of integrating channel attention mechanisms in U-shaped networks. Building on this concept, Buttar *et al.* (40) extended it by incorporating a U-Net++ (41) architecture with SEM modules, aiming to boost feature extraction capabilities. Partio *et al.* (42) proposed CloudCast, a U-Net-based model for total cloud cover nowcasting. Trained on five years of satellite data, it outperforms numerical weather prediction models and enhances short-term cloud forecasting.

With the growing emphasis on lightweight models, research has increasingly focused on balancing performance and efficiency, as seen in salient object detection (43; 44). In cloud segmentation, Li *et al.* proposed UCloudNet (27), which utilizes residual connections within U-Net to stabilize training in lightweight models. More recently, in 2025, Li *et al.* introduced DDUNet (28), incorporating weighted dilated convolution and a dynamic weight and bias generator to further enhance performance in compact architectures. Despite these advances, balancing segmentation accuracy and computational efficiency remains a persistent challenge. Our proposed SCANet aims to tackle this issue by leveraging lightweight architectures while maintaining high segmentation performance.

A.2 SCANET - BASIC BUILDING BLOCKS

The fundamental building blocks in computer vision deep learning-based tasks can be categorized as derivatives of three key modules: straight-forward structures, residual blocks (45), and inverted residual blocks (19). Fig. 4 (a) illustrates the straight-forward structure, which consists of a 3×3 convolution layer, a batch normalization layer, and a ReLU activation. This structure was widely used in early CNN models. Fig. 4 (b) presents the residual block architecture, which introduces a shortcut connection to facilitate the training of deep CNN networks. In contrast, Fig. 4 (c) depicts the inverted residual block, which first expands channels using a series of Conv2D-BatchNorm-ReLU6 layers. The features then pass through another set of similar layers, except that the standard Conv2D operation is replaced by depth-wise convolution (DWConv in Fig. 4), where the number of groups is set to the number of input channels, significantly reducing parameter count. Finally, a 1×1 convolution layer and a batch normalization layer reduce the channels back to the input dimension before applying element-wise addition.

Notably, the inverted residual block utilizes ReLU6 instead of ReLU, as ReLU6 caps the output value at 6, preventing excessively large activations. This bounded range helps maintain accuracy when performing inference in lower precision settings (e.g., FP16 mode).

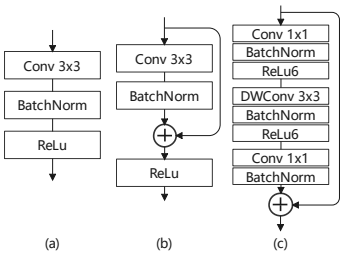

Table 4: Comparison of basic building blocks widely used in backbone networks. (a) shows the straightforward structure with a simple convolutional layer followed by batch normalization and activation, commonly employed in early CNN architectures. (b) illustrates the residual block (45) (c) demonstrates the inverted-residual block, designed to reduce parameters and computational cost through channel expansion and depth-wise separable convolutions (19).

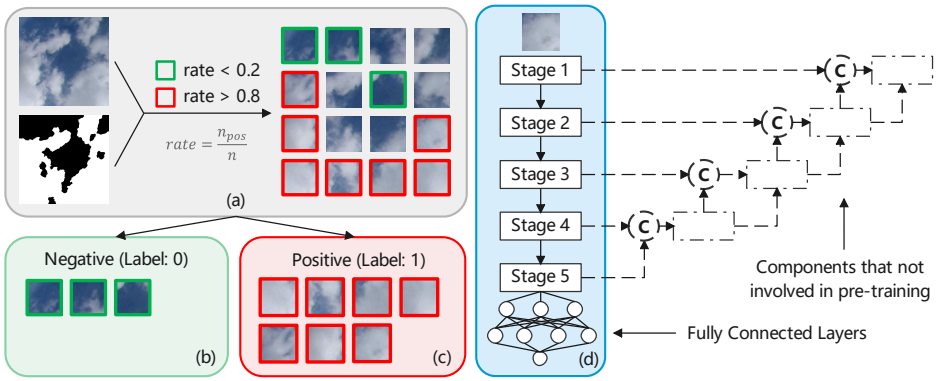

Figure 4: Schematic diagram of SWINySEG-based pre-training. (a) illustrates the positive and negative sample generation process. (b) indicates the negative samples. (c) is positive samples. (d) represents the modules involved in pre-training.

A.3 SWINYSEG-BASED PRE-TRAINING (SWPT)

Pre-training is widely used in computer vision tasks to improve performance and accelerate convergence, particularly for complex tasks such as semantic segmentation and object detection. Since training a model from scratch is computationally expensive, pre-training typically involves training the backbone on ImageNet (46) before fine-tuning on the target dataset. In SCANet and SCANet-large, we directly reuse pre-trained weights, as no modifications are made to the backbone. However, SCANet-lite introduces architectural changes that require pre-training from scratch. Given the high cost and time required for ImageNet pre-training, we propose an alternative strategy leveraging the SWINySEG dataset, as illustrated in Fig. 4.

Our approach involves iterating through the SWINySEG dataset, splitting each image into 16 patches, and assigning labels based on the proportion of cloud pixels in each patch:

$$\text{rate} = \frac{n_{\text{pos}}}{n} \tag{4}$$

where n_{pos} represents the number of cloud pixels (label 1), and n denotes the total number of pixels in the patch. If $\text{rate} > 0.8$, the patch is labeled as a positive sample (cloud); if $\text{rate} < 0.2$, it is labeled as a negative sample (sky). Patches with rate between 0.2 and 0.8 are ignored to ensure clear separation between classes. This threshold selection balances the number of positive and negative samples. This threshold selection helps to mitigate ambiguous regions, thus improving the clarity of the pre-training labels. By focusing only on well-defined cloud and sky regions, this strategy enhances the quality of feature representations learned by the model.

During pre-training, we remove all decoders and their connections to the backbone, replacing them with a fully connected layer to facilitate feature learning.

A.4 DATASET

We use the Singapore Whole Sky Nychthemeron Image SEGmentation Database (SWINySEG) as our training dataset, which consists of 6078 day-time and 690 night-time cloud images captured in Singapore using a calibrated camera. Following Zhang *et al.* (29), we split the dataset into training and testing sets with a 9:1 ratio. For evaluation, SCANet is tested on three subsets: day-time images (augmented SWIMSEG), night-time images (augmented SWINSEG), and the full SWINySEG dataset. Notably, SCANet is trained only once on the complete SWINySEG dataset.

A.5 IMPLEMENTATION DETAILS

We implement SCANet using PaddlePaddle and conduct training on a single NVIDIA Tesla V100-SXM2 16GB GPU. The model is trained for 100 epochs with a batch size of 16. We use the Adam optimizer with an initial learning rate of $1e-3$, $\beta_1 = 0.9$, $\beta_2 = 0.999$, and epsilon set to $1e-8$. The learning rate follows an exponential decay with a decay factor of $\gamma = 0.95$ after each epoch. Evaluation on the test set is conducted every 5 epochs to monitor performance.

For data augmentation, we apply only random horizontal and vertical flips after resizing the images to a resolution of 320×320. The augmented images are then scaled to the range $[0, 1]$ and normalized to have a mean of 0.5 across all three channels.

A.6 ADDITIONAL EXPERIMENT

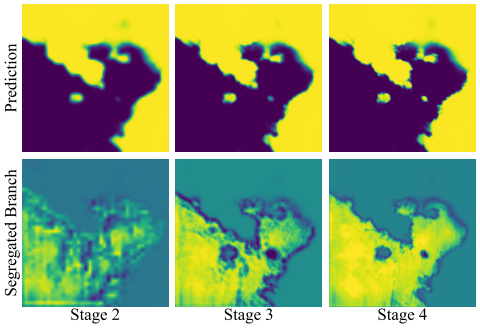

Figure 5: Visualization of SCAM output s_i (first row) and background mask m_i (second row)

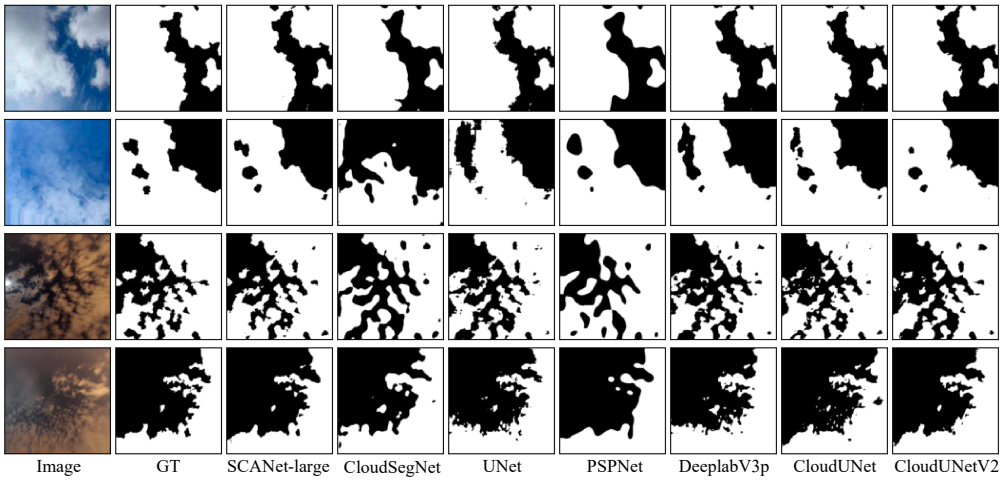

Figure 6: Additional Qualitative Experiments Comparing SCANet-large with State-of-the-Art Approaches on Daytime (rows 1–2) and Nighttime (rows 3–4) Images from the SWINySEG Dataset

We assess the effectiveness of the SCAM by visualizing its intermediate outputs, as illustrated in Fig. 5. The resolutions of the segregated branch output (background mask) from stages 2 to 4 are 40×40, 80×80, and 160×160, while the corresponding stage predictions have resolutions of 80×80, 160×160, and 320×320. At stage 2, the background mask is coarse but already outlines the general segmentation shape. By stage 3, the mask is refined with sharper boundaries, leveraging information from the previous stage. In the final stage, the background mask and stage prediction become well-defined, appearing nearly pure yellow in their combination. This confirms SCAM's effectiveness, particularly in enhancing feature utilization through element-wise operations (addition, multiplication, subtraction) and sigmoid activation—without requiring learnable parameters.

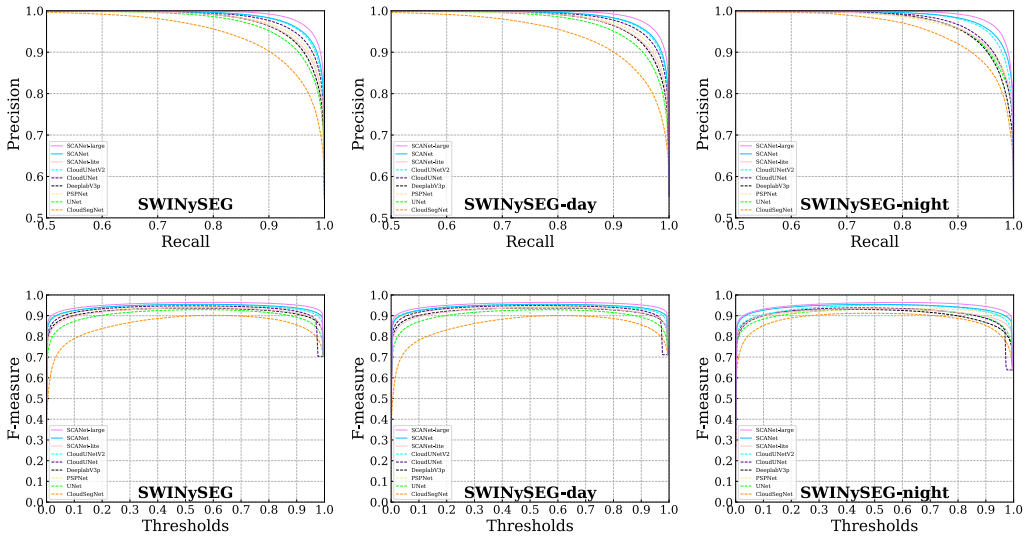

Figure 7: PR curves (first row) and F-measure curves (second row) on SWINySEG, SWINySEG-day, and SWINySEG-night dataset

We provide Fig. 6 to complement the qualitative experiment and Fig. 7 to supplement the quantitative experiment. In addition, to complement the ablation study, we provide Precision-Recall (PR) and F-measure curves for SCANet-large and six prior methods, including CloudSegNet (12), CloudUNet (22), CloudUNetv2 (23), U-Net (21), PSPNet (24), and DeepLabV3plus (25), as shown in Fig. 8. The PR curve illustrates the balance between precision and recall across various thresholds, while the F-measure curve highlights the model's performance at different thresholds, further validating SCANet's effectiveness in cloud segmentation. We also provided Fig. 9 showing visualizations of eight essential experiments (No. 1, 2, 3, 4, 5, 6, 8, 10).

A.7 Discussion

Ground-based sky/cloud segmentation extracts cloud structures from Earth-based observations, enabling cloud distribution visualization and supporting downstream meteorological applications, such as weather forecasting and anomaly detection. Deep learning significantly enhances accuracy and efficiency in this task. Early methods, like CloudSegNet (12), employ simple encoder-decoder architectures with convolution and max-pooling layers, ensuring computational efficiency but often falling short in accuracy for advanced meteorological analysis. To enhance performance, many approaches incorporate larger backbone networks or non-local attention mechanisms (34), which improve feature extraction but substantially increase computational complexity.

SCANet introduces a new strategy to optimize both accuracy and efficiency by utilizing lightweight backbone networks while refining decoder design through SCAM. SCAM employs two branches—Left and Right—to process features with high sky and cloud weights, respectively, leveraging prior decoder predictions. Their outputs are then combined to generate the final segmentation. As demonstrated in ablation experiments (No. 1, No. 2, No. 3) in Table 3, both branches contribute significantly to segmentation performance. This design ensures balanced consideration of sky and

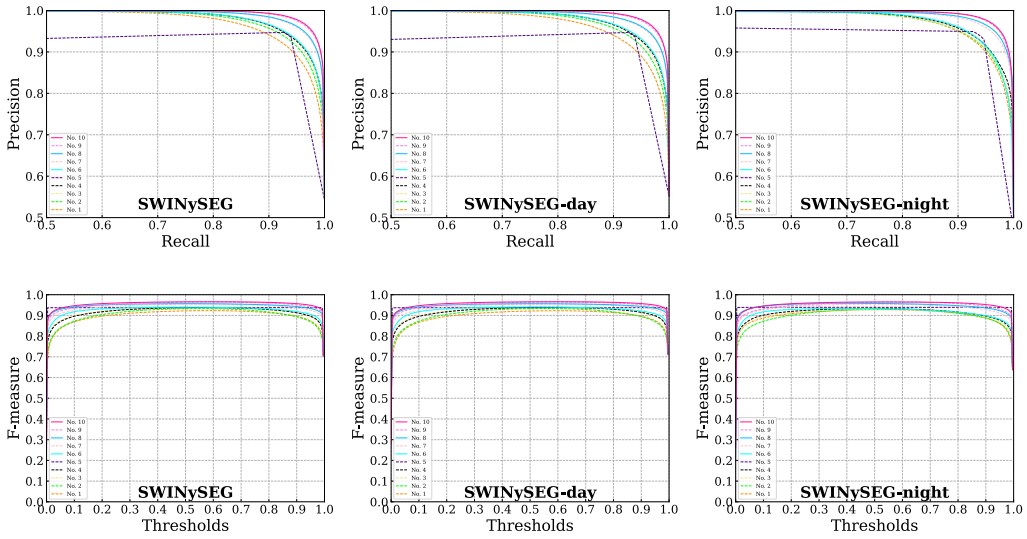

Figure 8: Illustration of PR curves (first column) and F-measure curves (second column) of ablation study in Table. 3 on SWINySEG, SWINySEG-day, and SWINySEG-night

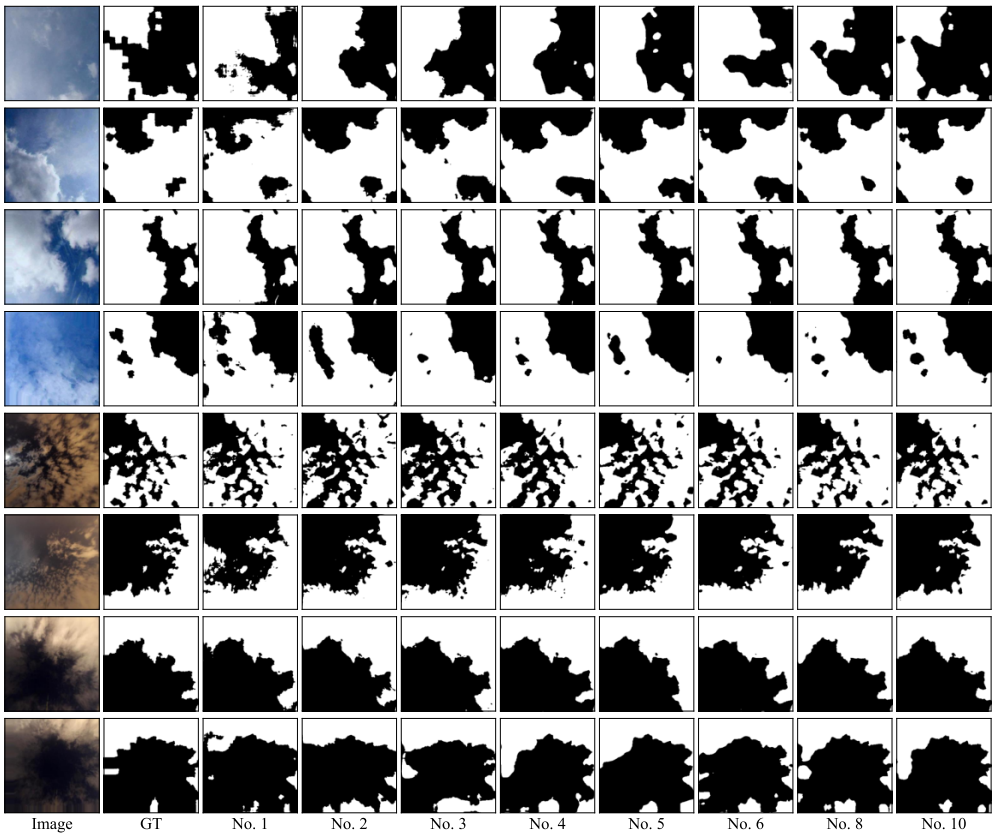

Figure 9: Qualitative visualization of ablation study on day-time and night-time images of SWINy-SEG dataset; This figure show the prediction maps of experiments No. 1, No. 2, No. 3, No. 4, No. 5, No. 6, No. 8, and No. 10 of Table. 3

cloud regions. Additionally, depth-wise convolution is used to minimize decoder parameters, effectively maintaining high segmentation accuracy while reducing computational cost.

A.8 Impact on Climate Change and ML-Driven Climate Solutions Community

Our SCANet model represents a significant step forward in applying machine learning to climate action, particularly in the field of atmospheric monitoring and extreme weather prediction. By improving cloud segmentation accuracy in real time, SCANet enhances climate models, facilitates solar irradiance forecasting, and supports disaster preparedness, making it a valuable tool for both climate mitigation and adaptation.

By improving the accuracy of cloud representation in atmospheric and climate models, SCANet enables a deeper understanding of cloud dynamics and albedo effects, which are essential for predicting global warming and refining climate simulations. Its ability to process cloud images at 1390 FPS ensures real-time analysis of sky conditions, making it highly effective for monitoring extreme weather events such as hurricanes, thunderstorms, and heatwaves. Additionally, SCANet plays a vital role in optimizing renewable energy systems by providing precise cloud segmentation for solar irradiance forecasting, which supports better integration of photovoltaic energy into power grids and reduces reliance on fossil fuels.

The model's lightweight design—requiring as few as 90K parameters in its smallest configuration—makes it highly efficient for deployment on edge devices, enabling cost-effective, real-time climate monitoring in remote and resource-limited areas. This accessibility facilitates localized monitoring of cloud cover, which complements satellite-based observations by adding high-resolution, ground-level data. Furthermore, SCANet contributes to disaster preparedness by enabling the rapid identification of cloud formations linked to severe weather, assisting in early warning systems and response planning.

Beyond weather prediction, SCANet contributes to climate science and carbon tracking, supporting research in cloud-albedo effects, CO_2 sequestration, and geoengineering interventions. Its scalability ensures adaptability across diverse geographies, helping bridge the gap between machine learning research and real-world climate applications. Its deployment also extends to research on geoengineering and carbon sequestration, where precise monitoring of cloud-albedo effects and atmospheric CO_2 levels can inform strategies for mitigating global warming.

By combining high accuracy, real-time efficiency, and deployability, SCANet exemplifies the potential of machine learning to enhance climate change mitigation and adaptation efforts on a global scale. We hope our work provides meaningful contributions to climate-focused machine learning research and inspires further exploration of efficient, real-time models to support climate change mitigation and adaptation efforts.