A Robust Attention Based Land Cover Classification Scheme for Corrupted Remote Sensing Images

Weipeng Shi, Wenhu Qin*, Zhonghua Yun, Tao Zhao, Chao Wu, Allshine Chen

***Abstract*—There is a significant effect of deep learning on the accuracy of land cover classification. Nonetheless, there is always a notable reduction of classification robustness in foggy conditions, which is commonly overlooked. Several challenging factors, like low image quality and occlusion, contribute to this. Instead of focusing solely on classification accuracy, we also explore the influence of attention mechanisms and multimodal fusion on classification robustness. ConvNeXt is adopted as the backbone. Furthermore, we propose Contextual Representation Enhancement Module (CREM) and Cross-Modal Fusion Module (CMFM) based on nonlocal operation. CREM possesses a large perceptive field to fuse local and global features, reducing side effects of the redundant noise. CMFM explores the relationship between multimodal inputs for information recalibration. Extensive ablation and comparison experiments were conducted on the corrupted ISPRS Potsdam and Vaihingen benchmark datasets to validate the proposed method. Compared to the reference model, our framework exhibits excellent accuracy and robustness in the task of land cover classification. Code will be available at https://github.com/bowenroom/Robust-land-cover-classification.**

***Index Terms*—Semantic segmentation, Attention mechanism, Robust deep learning, Remote sensing, Data fusion**

I. INTRODUCTION

Accurate and robust land cover classification (LCC, a.k.a. semantic segmentation) of remote sensing images (RSIs) is a prerequisite for a series of tasks, such as earth observation [1], crop growth monitoring [2], soil-permittivity estimation [3] and so on. Models with high robustness perform well on fog corrupted RSIs with less degradation compared to the clean. In general, LCC models, which apply deep learning with positive results, can be categorized into four types: FCNs [4], UNets [5], HRNets [6], and ViTs [7]. Despite the availability on clean data, they are not always reliable when corruptions exist.

Figure.1 illustrates some frequently encountered challenges in foggy conditions. It is evident from Figure.1(a) that there is a distinction in shapes of the cars. The fog interferes with visual representations of heterogeneous objects, resulting in a high degree of homogeneity in Figure.1(b) (d) (e). In Figure.1(c), a large proportion of the car is heavily obscured by trees and fog. It is tough to distinguish homogeneous cars in Figure.1(f) as a result of shadows and occlusion.

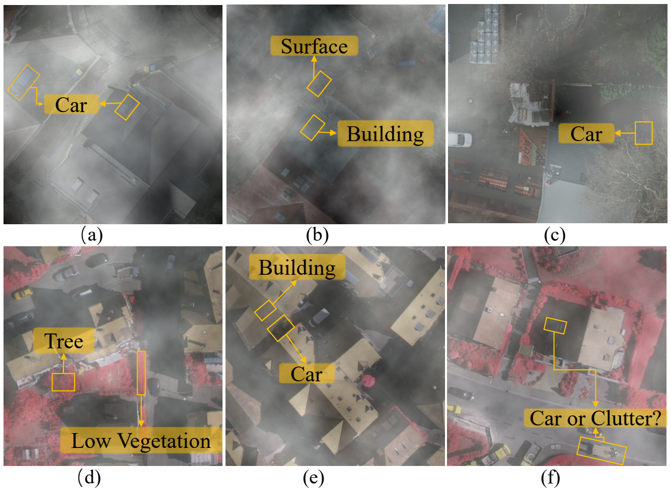
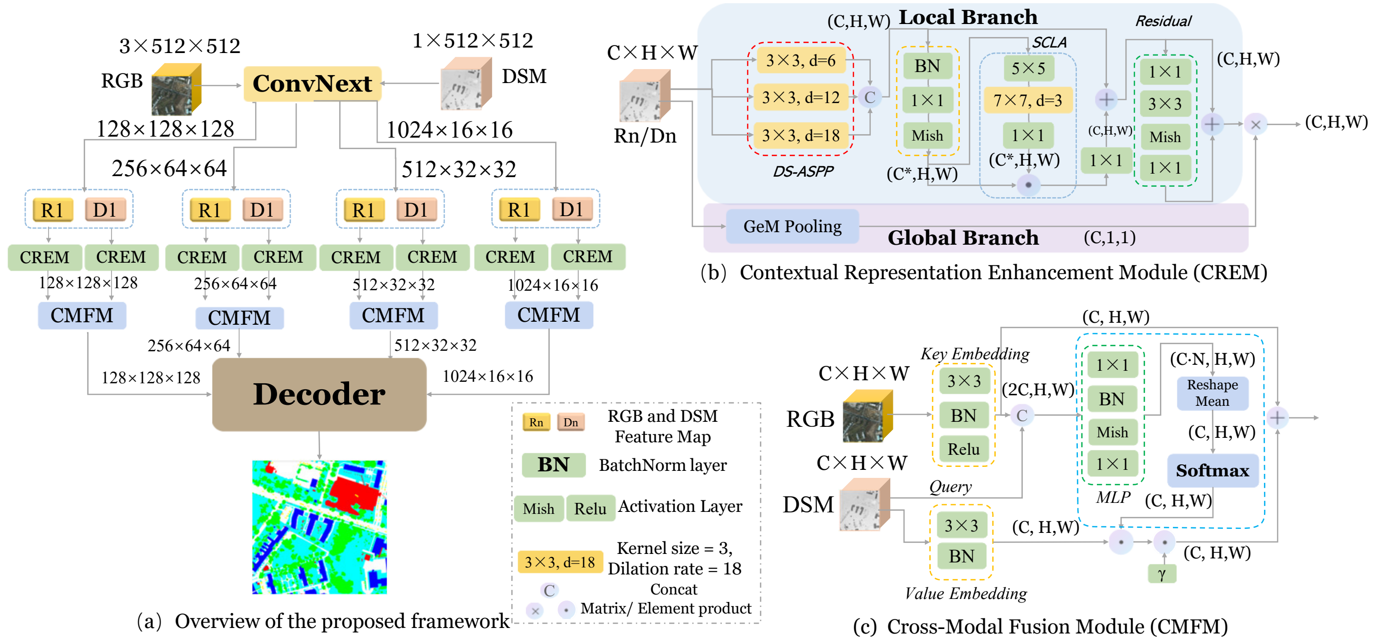


Fig. 1: Some challenging characteristcs of fog corrupted RSIs, such as diverse scales and shapes, heterogeneous homogeneity, occlusion as well as homogeneous heterogeneity.

To cope with the challenges encountered in fog, it is imperative to explore how to conduct LCC robustly. In view of the outstanding performance while handling images with poor quality, we adopt ConvNeXt [8] as the backbone. Meanwhile, we propose and incorporate additional two modules, CREM and CMFM, to improve the robustness. CREM consists of local and global branches. CREM is featured by a large receptive field which captures and integrates global as well as local features to reduce the redundant noise. Based on the non-local operation, we propose CMFM to recalibrate and explore the complementary interaction between multimodal inputs. In other words, our contributions can be summarized as the following:

- For land cover classification under fog, we develop an end-to-end model to improve the performance without significant degradation comparing to the clean condition.
- We propose CREM for feature extraction and CMFM for multimodal fusion based on the backbone ConvNeXt.
- A fog corrupted test set is generated for the robustness evaluation of each framework.

This work was supported by Key R&D Program of Jiangsu Province under Grant BE2019311, Jiangsu modern agricultural industry key technology innovation project under Grant CX(20)2013 and National key research and development program under Grant 2020YFB160070301. Weipeng Shi*(bowenroom1@gmail.com)*, Wenhu Qin**(corresponding author, qinwenhu@seu.edu.cn)*, Zhonghua Yun, Chao Wu are with School of Instrument Science and Engineering, Southeast University, Nanjing, 210096, China. Allshine Chen is with Health Sciences Center, University of Oklahoma, Oklahoma City, 73106, USA. Tao Zhao is with Institute of Intelligent Manufacturing, Nanjing Tech University, Nanjing 210009, China.



(a) Overview of the proposed framework

(b) Contextual Representation Enhancement Module (CREM)

(c) Cross-Modal Fusion Module (CMFM)

Fig. 2: Architecture of our model. It is composed of an encoder (ConvNeXt), CREM, CMFM and a decoder (UPerNet).

II. METHODOLOGY

A. Backbone and Decoder

According to the framework overview in Figure.2, both modalities are firstly assembled into ConvNeXt for encoding, which generates feature maps with four various resolutions. Low-level features are more exhaustive but contain redundant noise, whereas high-level features are enriched in semantic information but do not depict edges precisely. CREM is proposed to fuse local and global features, while CMFM is conducive to the extractions of complementary representations from multimodal inputs. According to the comprehensive comparison in [9], ResNeXt is capable of handling distorted images properly. In view of inherent inductive bias and translational invariance of CNNs, ConvNeXt, which follows the ResNeXt design scheme of using group convolution to extend model width, is more robust than ViT. Furthermore, equipment of the large kernel convolution provides a large and global receptive field. Therefore, we deploy ConvNeXt as the backbone, which allows the network to effectively cope with low-quality RSIs. UPerNet is selected as the decoder, which implements a top-down lightweight structure accompanied by a feature pyramid network and a pyramid pooling module to exploit hierarchical levels of representations. In the context of complex remote sensing scenarios, the combination of parsed visual attributes improves classification robustness.

B. Context Representation Enhancement Module

If corrupted multimodal inputs (denoted as $\boldsymbol{X}$, $X \in R^{C\times H\times W}$) collected under fog are fed directly to the fusion module, the network will be subject to an increased level of redundancy and noise. To reduce interference of the noise on robustness, we propose CREM, illustrated in Figure.2, to boost the capability of obtaining useful information. It is composed of local and global branches, which strengthen the complementarity and compatibility between multiple modalities. The acquired global representation is generally more tolerant of variations in illumination and viewing angles in RSIs, while the local representation is more sensitive to edge details, geometry and texture. Calibration fusion filters the noise information.

The local branch consists of a Depthwise Separable Astrous Spatial Pyramid Pooling (DS-ASPP), a Spatial and Channel Large kernel Attention (SCLA) and a residual module. DS-ASPP can efficiently cope with the scale diversity of instances in RSIs with a marginal cost. Depthwise Separable convolution [10] is comprised of depthwise and pointwise convolutions, where the depthwise performs a separate spatial convolution for each channel and the pointwise combines the generated outputs. After concatenating output from DS-ASPP (denoted as $\boldsymbol{X_{aspp}}$, $X_{aspp} \in R^{C\times H\times W}$), the embeded result ($\boldsymbol{X^*_{aspp}}$, $X^*_{aspp} \in R^{C\times H\times W}$) is acquired by a 1×1 convolution ($\boldsymbol{K_1^{1\times1}}$). Self attention mechanism ignores the adaptive characteristics of each channel. We get the attention map ($\boldsymbol{X_{scla}}$, $X_{scla} \in R^{C^*\times H\times W}$) after SCLA, which consists of three types of convolution($\boldsymbol{K^{1\times1}}$, $\boldsymbol{K^{DW\_D}}$, $\boldsymbol{K^{DW}}$ denote 1×1 convolution, depth-wise dilation convolution, depth-wise convolution respectively). A 1×1 convolution is applied for channel alignment. The fused result ($\boldsymbol{\hat{X}_{scla}}$, $\hat{X}_{scla} \in R^{C\times H\times W}$) can model the importance of each channel and location in feature maps, thus capturing the long-range dependencies in RSIs well. Summed with the original input, we use a residual module to reduce the information redundancy. $\boldsymbol{Y_{glocal}}$ ($Y \in R^{C\times1\times1}$) is the feature map after generalized-mean pooling layer (GeM Pooling [11]), which can generate an image descriptor about spatial distribution of distinction influence. p_n is the pooling parameter, controlling amount of the response correspondence. $\otimes$ denotes the element-wise product. Therefore, we could explore the complementary context through matrix product fusion of two branches. Formula details are displayed as the following:

$$\mathbf{X}^{*}_{aspp} = \mathbf{K_1}^{1\times 1} \cdot \mathbf{X}_{aspp} \tag{1}$$

$$X_{scla} = \mathbf{K_2^{1\times 1}} \cdot \mathbf{K^{DW\_D}} \cdot \mathbf{K^{DW}} \cdot \mathbf{X}^{*}_{aspp} \tag{2}$$

$$\hat{X}_{scla} = \mathbf{K_3^{1\times 1}} \cdot (\mathbf{X}^{*}_{aspp} \otimes X_{scla}) + \mathbf{X}_{aspp} \tag{3}$$

$$\mathbf{Y_{local}} = \mathbf{F}(\hat{X}_{scla}) + \hat{X}_{scla} \tag{4}$$

$$\mathbf{Y_{global}} = \left[\mathrm{y}_1^{(g)} \ldots \mathrm{y}_n^{(g)} \ldots \mathrm{y}_C^{(g)}\right]^{\top}, \mathrm{y}_n^{(g)} = \left(\tfrac{1}{|\mathcal{X}_n|} \textstyle\sum_{x\in\mathcal{X}_n} x^{p_n}\right)^{\frac{1}{p_n}} \tag{5}$$

C. Cross-Modal Fusion Module

Conventional early-fusion algorithm fuse multimodal data ineffectively, since it can not acquire interactive characteristics and relationship between modalities adequately. Self-attention scheme efficiently activates the feature map which is associated with various spatial locations of inputs. Figure.2 illustrates CMFM. Rather than purely relying on self-attention, we combine contextual information from RGB input ($\boldsymbol{X_{RGB}}$, $X_{RGB} \in R^{C\times H\times W}$) with sparse, high-quality height information from DSM ($\boldsymbol{X_{DSM}}$, $X_{DSM} \in R^{C\times H\times W}$). Consequently, the relationship between adjacent context can be effectively exploited, resulting in an aggregated output.

The key embedding is based on 3×3 depth-wise convolution W_k^D, which contextualizes the key representation over spatial locations. We then acquire the attention matrix $\boldsymbol{Y_{attn}}$ by passing concatenated feature maps through MLP. The embeded key feature map $W_k^D X_{RGB}$ acts as the static context, which guides the key learning points. γ is the learnable scalar initiallizing from 0, which controls the weight of dynamic contextual representation. Utilizing element-sum, we explore complementary interactions of multimodal inputs. This process is formulated as:

$$\mathbf{Y_{attn}} = W_{mlp} \cdot [Concat(W_k^D X_{RGB}, X_{DSM})] \tag{6}$$

$$\mathbf{Y_{output}} = W_k^D X_{RGB} + \gamma \cdot [Softmax(Y_{attn}) \cdot W_q X_{DSM}] \tag{7}$$

III. EXPERIMENTS AND ANALYSIS

A. Dataset

ISPRS[1] offers Potsdam and Vaihingen LCC datasets for the experiment. Potsdam has a resolution and ground sampling distance of 6000 × 6000 pixels and 5cm respectively. Corresponding parameters for Vaihingen are 2500 × 2500 pixels and 9cm. Image in Potsdam indexed **4_12** was removed as a result of the error existing in the annotation. In accordance with the ratio of 0.8, 0.1, and 0.1, each dataset is divided into a training set, a validation set, and a test set. Building, Car, Vegetation, Tree, Surface as well as Clutter make up the entire dataset. Figure.3 demonstrates the proportion of different classes in both aerial datasets. In Figure.3, it is clear that buildings accompanied by the impervious surface account for more than half of all pixels. The percentage of Car is considerably inferior to the other items. Clutter in Potsdam is 7.4 times greater than in Vaihingen. This phenomenon arises from the complicated attribute of city scenes.

Data augmentation techniques, including random rotation, reflection padding, random adjustment of basic attributes and so on, are applied to reduce overfitting and class imbalance issues. The original image, which is captured in normal conditions, is applied for model training and validation. To evaluate the robustness performance in various circumstances, we create the fog corrupted dataset based on diamond-square algorithm [12]. Figure.4 displays several conditions with different degrees of fog corruptions.

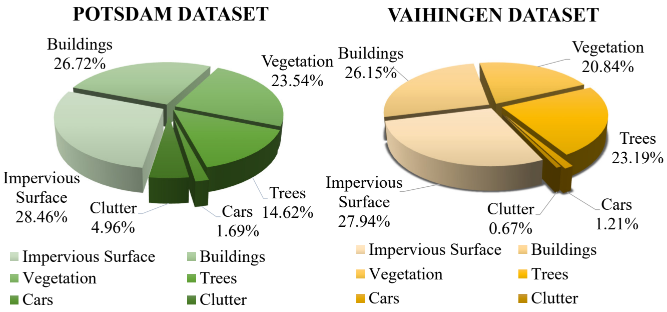


Fig. 3: Class imbalance exists in both datasets.

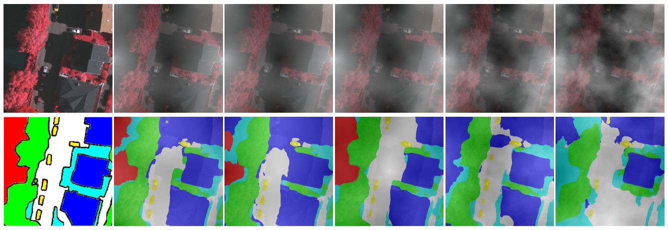


Fig. 4: The first column displays a clean Vaihingen image with the ground truth. Fog concentration gradually increases from left to right and the second row corresponds to LCC result.

B. Metrics

Metrics utilized to evaluate the performance are F1 score ($F1_{score}$), Overall Accuracy (**OA**). Corruption Degradation (**CD**). We define $\boldsymbol{CD}$ [12] and Degradation $\boldsymbol{D}$ as the following:

$$D = 1 - F1_{score} \tag{8}$$

$$CD^f = \frac{\widetilde{D^f}}{\widetilde{D^{ref}}} \times 100\% \tag{9}$$

Specifically speaking, OA as well as $F1_{score}$ are to evaluate LCC accuracy and CD assesses LCC robustness in terms of various fog corruptions. Larger values of OA and $F1_{score}$ indicate the better performance in LCC accuracy, whereas larger CD value represents a lower robustness with higher LCC performance degradation. *ref* and *f* denote the reference and selected models. D represents the absolute robustness degradation where $\widetilde{D^f}$ indicates the average degradation over various levels of fog corruption.

[1] Potsdam and Vaihingen datasets can be acquired from *https://www.isprs.org/education/benchmarks/UrbanSemLab/2d-sem-label-potsdam.aspx* and *https://www.isprs.org/education/benchmarks/UrbanSemLab/2d-sem-label-vaihingen.aspx* (*Accessed in June, 2022*)

C. Implementation Details

Training and inference are performed on one RTX 3090. To fit into GPU memory, we slice the high-resolution image into 512x512 pixels. We assemble RGB and DSM into one image by means of RGB channels and alpha channel. Mish activation is also adopted in Figure.2 [13]. Dilation rates in DS-ASPP are [6,12,18]. C^{*}, N, p_n in Figure.2 and Equation.(5) are set as 300, 3, 32 respectively. With blocks of [3, 3, 27, 3] for each stage, ConvNeXt (**Base**) pretrained on ImageNet is adopted as the backbone and we take UperNet as the decoder. The learning rate is set initially as 1e-4 using a poly learning rate schedule with a power of 1. Weight decay is 5e-2. To reduce the memory consumption, we train each model in 40k iterations with the mixed precision training strategy.

TABLE I: Ablation study of CREM on Potsdam dataset

Index	Backbone	CREM			mFscore(%)↑		CD(%) ↓
		DS-ASPP	SCLA	GEM-P	Clean	Fog	
A	*ConvNeXt*				82.82	57.13	100.00
B	*Segformer*				79.74	55.94	102.78
C	*ConvNeXt*	✓			84.90	62.74	86.91
D	*ConvNeXt*		✓		85.12	63.07	86.14
E	*ConvNeXt*			✓	83.10	61.53	89.74
F	*ConvNeXt*	✓	✓	✓	**86.01**	**64.63**	**82.51**

TABLE II: Ablation study of CMFM on Potsdam dataset

Index	Block design	GFLOPs	mFscore(%)↑		CD(%) ↓
			Clean	Fog	
A	*Baseline*	2.78	80.45	50.19	100.00
B	*+SNL*	3.29	82.59	57.99	84.34
C	*+SNL +Concat*	3.29	83.08	59.06	82.19
D	*+SNL+FM*	4.14	84.64	63.13	74.02
E	*+QE+Concat+FM*	4.98	**86.21**	63.98	72.31
F	*+ CMFM*	4.15	86.01	**64.63**	**71.01**

D. Results and Analysis

Model performance is evaluated on both *Clean* and *Fog* corrupted datasets. *mFscore* represents calculating the mean $F1_{score}$ across five severity levels of various categories. Higher values of *OA* and *mFscore* signify a strong performance in accurate classfication. Comparing to the reference, lower *CD* indicates a high degree of robustness with low degradation from clean conditions.

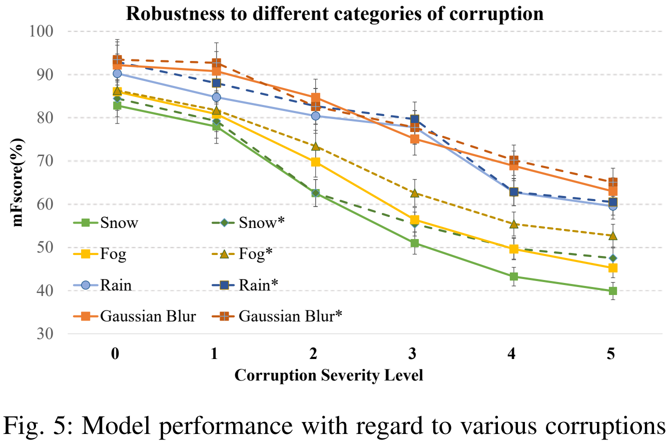


Fig. 5: Model performance with regard to various corruptions.

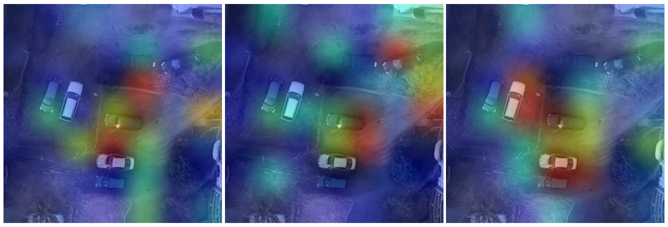

(a) + CREM (b) + CMFM (c) + CREM + CMFM

Fig. 6: The generated heat map associated with Car.

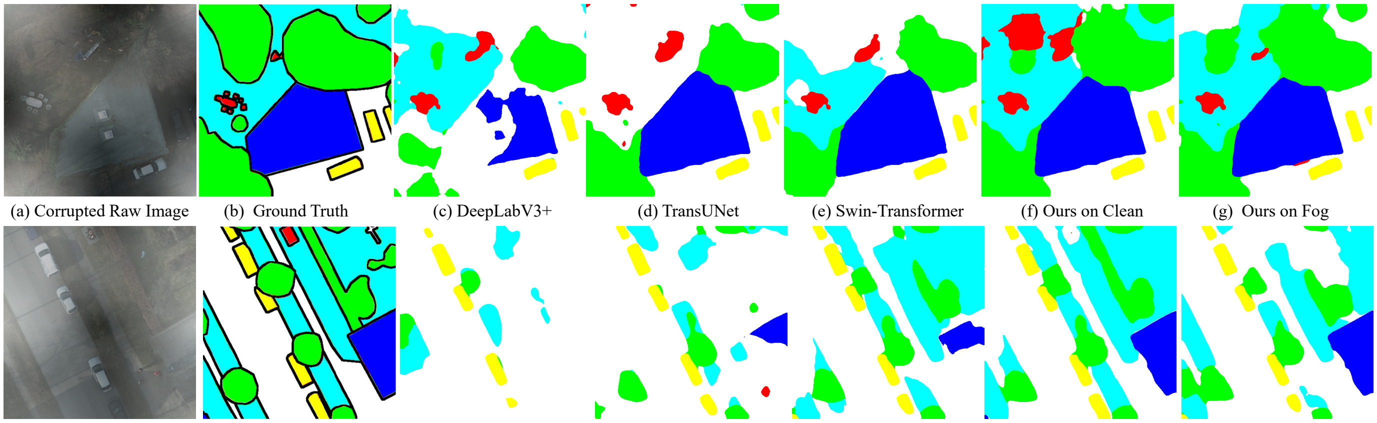
1) Ablation Study: This section is to compare the effect of proposed constituent. We integrate each part into the backbone with A as the reference. According to A and B in Table.I, when performing downstream semantic segmentation tasks, ConvNeXt achieves 3.42% better classification accuracy and 2.78% better robustness than Segformer [18] as SOTA. The cross comparison between A with C, D, and E reveals that SCLA contributes most in terms of improving classification accuracy and robustness (13.86% reduction in CD). In summary, the model that incorporates all components demonstrates the superior performance (a decrease in CD of 17.49% and a 3.19% improvement in mFscore compared to the reference). Table.II compares each part of CMFM. +*SNL* means the Simplifed Nonlocal block, which is derived from GCNet [19]. The blue dashed box in Figure.2 is the Fusion module(*FM*). *Concat* simply concatenates multimodal inputs. It can be concluded from the combination of A and B in Table.II that SNL effectively improves the accuracy and robustness in both clean and foggy conditions. Based on the comparison of *BC* and *BD*, FM improves the classification accuracy and robustness by 4.07% and 8.17% respectively in the corrupted environment. The additional query embedding(+*QE*) of dsm makes E different from F. Despite the 0.2% improvement under clean, there is an extra 0.83 GLOPs and no reduction in CD. CMFM achieves a balance between efficiency, accuracy, and robustness.

Figure.5 illustrates the performance with regard to various corruptions in RSIs. $*$ denotes utilizing the corrupted as well as clean data in the training stage. It is beneficial to enhance robustness while exposing model to both scenes during the training stage. Additionally, our model performs better when coping with Gaussian blur, whereas the performance degrades the most in snow conditions. We also generate the feature map with regard to the class Car in Figure.6. Surrounding pixels can be well activated. With the aid of the effective multimodal fusion, our model is characterized with strong robustness due to the improved generalization of capturing long-range dependencies.

2) Comparison Study: Table.III shows result of the comparison study. Our model performs best for Car among all classes. TRM, which is the second best, is outperformed by 6.7% in mFscore under fog and 2.26% in CD value. We also observe that Transformer-based methods tend to perform better under clean, while the performance degradation is more severe in foggy conditions. This is attributed to the fact that CNN methods possess better induction bias, while ViTs require more data as pre-training to improve generalization. Our model has better performance in terms of accuracy and robustness

TABLE III: COMPARISON STUDY ON THE VARIANTS OF POTSDAM TEST SET

Method	Imp Surf *			Building			Low Veg *			Tree			Car			mFscore(%)			OA(%)		
	Clean↑	Fog↑	CD↓	Clean↑	Fog↑	CD↓	Clean↑	Fog↑	CD↓	Clean↑	Fog↑	CD↓	Clean↑	Fog↑	CD↓	Clean↑	Fog↑	mCD↓	Clean↑	Fog↑	CD↓
FCN [4]	82.20	52.90	100.00	85.49	55.14	100.00	68.20	30.21	100.00	70.04	49.12	100.00	67.49	28.85	100.00	74.68	43.24	100.00	79.55	51.06	100.00
UNet [5]	83.18	55.99	93.44	88.99	66.47	74.74	70.33	31.32	98.41	72.63	53.12	92.14	70.93	41.99	81.53	77.21	49.78	88.05	81.81	52.27	86.98
DeepLabV3+ [14]	86.72	58.39	88.34	88.49	69.48	68.03	73.52	37.68	89.30	77.93	55.61	87.24	72.42	52.46	66.82	79.82	54.72	79.95	83.10	59.45	78.27
TransUNet [15]	87.10	57.02	91.25	89.11	65.81	76.21	79.06	38.99	87.42	81.02	55.70	87.07	80.19	52.73	66.44	83.30	54.05	81.68	83.75	56.61	79.76
Swin-Transformer [7]	87.13	59.59	85.80	91.35	74.61	56.60	79.31	**57.63**	**60.71**	84.19	60.97	76.71	82.98	55.54	62.49	84.99	61.67	68.46	85.14	60.09	64.99
Swin-Unet [16]	87.67	58.41	88.30	93.56	76.81	51.69	80.69	57.27	61.23	84.32	65.71	67.39	84.34	55.95	61.91	86.12	62.83	66.11	85.83	60.27	61.67
TRM [17]	90.24	60.43	84.01	94.59	78.65	47.59	82.39	55.15	64.26	**89.94**	**72.98**	**53.11**	86.80	53.50	65.35	88.79	64.14	62.87	88.04	62.19	58.64
Ours	**92.95**	**65.05**	**74.20**	**96.96**	**84.12**	**35.40**	**87.38**	42.36	82.59	89.42	71.07	56.86	**96.36**	**61.59**	**53.98**	**92.61**	**70.84**	**60.61**	**90.91**	**66.08**	**47.77**



(a) Corrupted Raw Image (b) Ground Truth (c) DeepLabV3+ (d) TransUNet (e) Swin-Transformer (f) Ours on Clean (g) Ours on Fog

(a) Corrupted Raw Image (b) Ground Truth (c) DeepLabV3+ (d) TransUNet (e) Swin-Transformer (f) Ours on Clean (g) Ours on Fog

Fig. 7: Qualitative comparisons of several approaches for semantic segmentation of RSIs. The original image is corrupted by the fog of severity level three.

compared with other SOTAs.

IV. CONCLUSION

In this letter, we design a robust network for LCC in foggy conditions. Based on ConveNext, we propose CREM and CMFM based on attention mechanism and multimodal fusion to enhance robustness. The ablation study and comparison experiment demonstrate that our model is robust to fog corrupted RSIs, maintaining a balance between accuracy and robustness. It is our hope that this inspiring letter will provide an opportunity for researchers to explore the robustness in the field of remote sensing and earth observation more generally.

REFERENCES

[1] M. Schmitt, L. H. Hughes, C. Qiu, and X. Zhu, "Sen12ms - a curated dataset of georeferenced multi-spectral sentinel-1/2 imagery for deep learning and data fusion," *ArXiv*, vol. abs/1906.07789, 2019.

[2] D. Mandal, V. Kumar, V. Kumar, D. Ratha, S. Dey, A. Bhattacharya, J. M. Lopez-Sanchez, H. Mcnairn, and Y. S. Rao, "Dual polarimetric radar vegetation index for crop growth monitoring using sentinel-1 sar data," *Remote Sensing of Environment*, vol. 247, p. 111954, 2020.

[3] R. Hänsch, T. Jagdhuber, and B. Fersch, "Soil-permittivity estimation under grassland using machine-learning and polarimetric decomposition techniques," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 59, pp. 2877–2887, 2021.

[4] E. Shelhamer, J. Long, and T. Darrell, "Fully Convolutional Networks for Semantic Segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 4, pp. 640–651, 2017.

[5] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional Networks for Biomedical Image Segmentation," *arXiv:1505.04597 [cs]*, 2015.

[6] J. Wang, K. Sun, T. Cheng, B. Jiang, C. Deng, Y. Zhao, D. Liu, Y. Mu, M. Tan, X. Wang, W. Liu, and B. Xiao, "Deep High-Resolution Representation Learning for Visual Recognition," *arXiv:1908.07919 [cs]*, 2020.

[7] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin Transformer: Hierarchical Vision Transformer using Shifted Windows," *arXiv:2103.14030 [cs]*, Aug. 2021.

[8] Z. Liu, H. Mao, C.-Y. Wu, C. Feichtenhofer, T. Darrell, and S. Xie, "A ConvNet for the 2020s," *arXiv:2201.03545 [cs]*, Jan. 2022.

[9] S. Tang, R. Gong, Y. Wang, A. Liu, J. Wang, X. Chen, F. Yu, X. Liu, D. Song, A. Yuille, P. H. S. Torr, and D. Tao, "RobustART: Benchmarking Robustness on Architecture Design and Training Techniques," *arXiv:2109.05211 [cs]*, 2021.

[10] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, "MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications," *arXiv:1704.04861 [cs]*, 2017.

[11] M. Yang, D. He, M. Fan, B. Shi, X. Xue, F. Li, E. Ding, and J. Huang, "DOLG: Single-Stage Image Retrieval with Deep Orthogonal Fusion of Local and Global Features," *arXiv:2108.02927 [cs]*, Aug. 2021.

[12] C. Kamann and C. Rother, "Benchmarking the Robustness of Semantic Segmentation Models with Respect to Common Corruptions," *Int. J. Comput. Vis.*, vol. 129, no. 2, pp. 462–483, Feb. 2021.

[13] D. Misra, "Mish: A self regularized non-monotonic activation function," in *BMVC*, 2020.

[14] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation," *arXiv:1802.02611 [cs]*, Aug. 2018.

[15] J. Chen, Y. Lu, Q. Yu, X. Luo, E. Adeli, Y. Wang, L. Lu, A. L. Yuille, and Y. Zhou, "Transunet: Transformers make strong encoders for medical image segmentation," *ArXiv*, vol. abs/2102.04306, 2021.

[16] H. Cao, Y. Wang, J. Chen, D. Jiang, X. Zhang, Q. Tian, and M. Wang, "Swin-unet: Unet-like pure transformer for medical image segmentation," *ArXiv*, vol. abs/2105.05537, 2021.

[17] W. Shi, W. Qin, Z. Yun, A. Chen, K. Huang, and T. Zhao, "Land Cover Classification in Foggy Conditions: Toward Robust Models," *IEEE Geoscience and Remote Sensing Letters*, vol. 19, pp. 1–5, 2022.

[18] E. Xie, W. Wang, Z. Yu, A. Anandkumar, J. M. Alvarez, and P. Luo, "SegFormer: Simple and Efficient Design for Semantic Segmentation with Transformers," *arXiv:2105.15203 [cs]*, Jun. 2021.

[19] Y. Cao, J. Xu, S. Lin, F. Wei, and H. Hu, "GCNet: Non-local Networks Meet Squeeze-Excitation Networks and Beyond," *arXiv:1904.11492 [cs]*, Apr. 2019.