Land Cover Classification in Foggy Conditions: Toward Robust Models

Weipeng Shi, Wenhu Qin, Zhonghua Yun, Allshine Chen, Kai Huang, and Tao Zhao

***Abstract*— Robust semantic labeling of high-resolution remote sensing images (RSIs) in foggy conditions is crucial for automatic monitoring of land covers. This remains a challenging task owing to the low interclass differentiation yet high intraclass variance and geometric size diversity. Although conventional convolutional neural networks (CNNs) have demonstrated state-of-the-art (SOTA) performance in semantic segmentation, most networks are primarily concerned with standard accuracy, while the influence on robustness is rarely explored. This letter proposes a reliable framework which is evaluated across various severity levels of fog corruptions. Utilizing HRNet as the backbone to maintain high-resolution representations, we develop a multimodal fusion module (MMF) to exploit the complementary information of lidar and multispectral data. Based on the evaluation experiment on fog corrupted datasets, our model demonstrates promising performance with an average mean Intersection over Union (mIoU) on the clean along with the corrupted datasets exceeding 80% and 56%, respectively.**

***Index Terms*— Attention mechanism, data fusion, remote sensing, robust deep learning, semantic segmentation.**

I. Introduction

ROBUST semantic segmentation of corrupted remote sensing images (RSIs) means that computer vision models can perform comparatively accurate land cover classification in foggy conditions with minimal performance degradation compared to clean conditions. Fog, which impairs significantly the performance of mineral domain mapping [1], road extraction [2], 3-D reconstruction [3], change detection [4], is often encountered in RSIs. Deep learning based semantic segmentation models can be summarized into four categories, which are fully convolutional networks (FCNs) [5], UNets [6], HRNets [7] and Visual Transformers [8]. FCNs attain low-resolution representations by serially linking high-to-low convolutions. UNets restore the high-resolution representations through upsampling. HRNets learn ample semantic features with convolution layers in parallel through the multiresolution fusion units. Visual Transformers have achieved state-of-the-art (SOTA) results in several vision tasks owing to the powerful modeling capability. Segmentation transformer (SETR) [8] proposes a multilevel feature fusion module for fine-grained classification. Segformer [9] eliminates the time-consuming position encoding module in favor of a lightweight multilayer perceptron (MLP). Additionally, Zheng *et al.* [10] possessed both generalization and discrimination abilities through joint learning. For the semantic consistency, Ning *et al.* [11] considered pairwise and intramodal relationships simultaneously, along with nonpaired intermodal relationships. Zheng *et al.* [12] proposed deep scene representation for the context invariance of convolutional neural networks (CNNs).

Fig. 1. Challenges associated with robust classification of land covers in fog-covered regions.

Nonetheless, most semantic segmentation models are trained and validated on natural scene images (NSIs). Previous studies have rarely demonstrated the model's robustness and effectiveness in foggy conditions. Distinct from NSIs, RSIs are of poor quality, which is attributed to the susceptible data acquisition process, particularly when fog exists. As illustrated in Fig. 1, corrupted RSIs pose challenges to robustness, most notably in the following areas.

1) *Intraclass Heterogeneity and Interclass Homogeneity:* Models usually classify objects of the same type into different categories when the shapes and materials differ significantly. In contrast, objects of different classes are occasionally regarded to be identical incorrectly if they share similarities in the appearance. These are intraclass heterogeneity and interclass homogeneity issues.
2) *Occlusion and Geometric Size Diversity:* Complex 3-D interactions exist in RSIs, such as cars parked under the tree shadow or in haze. The accurate classification of spatial relationships cannot be guaranteed by optical cameras alone. Occlusion may lead to large variations in shape and size of objects belonging to the same class.

To alleviate the influence of aforementioned factors associated with remote sensing in foggy circumstances, a framework

Manuscript received 6 April 2022; revised 15 June 2022; accepted 25 June 2022. Date of publication 1 July 2022; date of current version 19 July 2022. This work was supported in part by the Key Research and Development Program of Jiangsu Province under Grant BE2019311, in part by the Jiangsu Modern Agricultural Industry Key Technology Innovation Project under Grant CX(20)2013, and in part by the National Key Research and Development Program under Grant 2020YFB160070301. *(Corresponding author: Wenhu Qin.)*

Weipeng Shi, Wenhu Qin, Zhonghua Yun, Kai Huang, and Tao Zhao are with the School of Instrument Science and Engineering, Southeast University, Nanjing 210096, China (e-mail: bowenroom1@gmail.com; qinwenhu@seu.edu.cn).

Allshine Chen is with the Health Sciences Center, The University of Oklahoma, Oklahoma City, OK 73106 USA.

Digital Object Identifier 10.1109/LGRS.2022.3187779

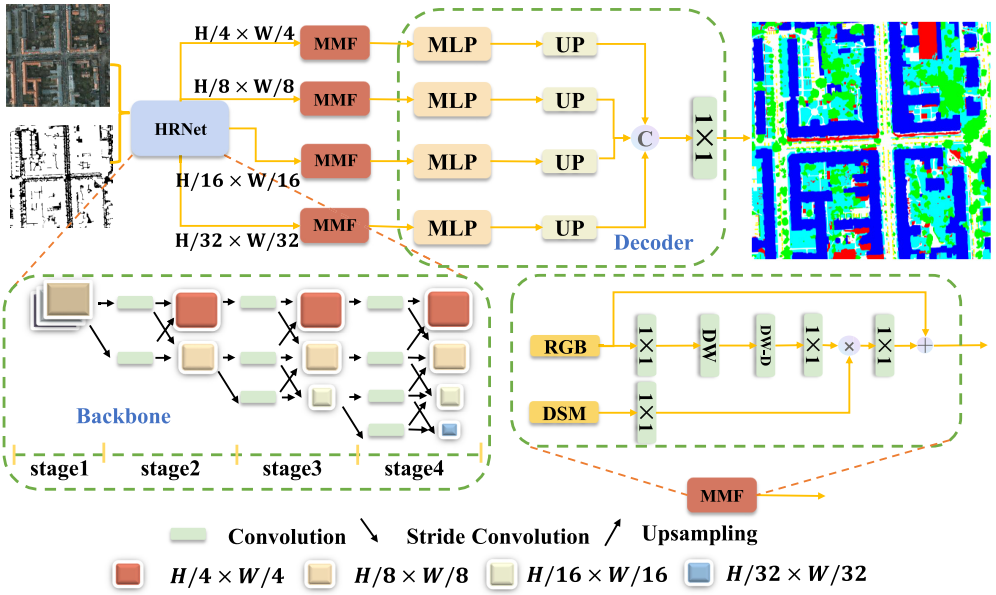

Fig. 2. Overall architecture. Our model consists of an encoder (HRNet), an MMF and a decoder. In the structure of MMF, **DW** and **DW-D** stand for depthwise convolution and depthwise dilation convolution, whose kernel sizes are 5 and 7 (dilation rate = 3), respectively.

for robust land cover classification is proposed. HRNet can maintain a consistent high-resolution representation throughout the network. Due to the adaptability to various scales and shapes, we employ HRNet as the backbone to tackle the intraclass heterogeneity issue. To address occlusion and interclass homogeneity issues, the proposed multimodal fusion module (MMF) investigates the complementary feature maps from the optical input and digital surface model (DSM). Our contributions can be summarized as follows.

1) We design an end-to-end semantic segmentation model which is robust for land cover classification in foggy conditions.
2) We incorporate MMF into the network leveraging HRNet as a backbone. MMF captures spatial and channel adaptability by utilizing the multimodal remote sensing data efficiently.
3) We conduct experiments on the fog corrupted International Society for Photogrammetry and Remote Sensing (ISPRS) test set to evaluate the effectiveness of our framework.

II. Methodology

A. Backbone

Aerial images captured in fog are of poor quality and intraclass heterogeneity is frequently encountered in the process of feature extraction with downsampling. HRNet is adopted as the backbone since it preserves high-resolution representations. The output can acquire rich semantic features without relying on upsampling to recover the high-resolution feature map from the low-resolution. As illustrated in Fig. 2, HRNet is composed of four separate branches with different resolutions (1/4, 1/8, 1/16, 1/32 of the original). Each branch is further split into four stages, and the output channel numbers are $C, 2C, 4C, 8C$. Multiresolution fusion blocks, existing between each stage, facilitate information flowing across branches. They are denoted as crossed lines (3×3 stride convolution and upsampling layer).

All these branches share information with each other on a variety of scales. High-resolution branches focus on spatial details, while lower branches highlight semantic features. HRNet can encode local and spatial information through multiresolution fusion, acquiring discriminative representations to appropriately classify different objects with the same semantic labeling. This enhances the robustness to intraclass heterogeneity significantly.

B. Multimodal Fusion Module

Adjacent objects of various classes may exhibit a similar appearance, but owing to different heights, they appear distinctly in DSM, which is advantageous for landscape modeling and visualization. Based on this attribute, we propose MMF for multimodal data fusion, which generalizes well when dealing with occlusion and interclass homogeneity. We perform grayscale mapping for DSM, which has three channels after the channel replication. According to the conclusion from GCNet [13], the gap between feature maps generated by different query locations is narrow, so we design the query independent structure. Originally, the self-attention structure acquires solely the spatial adaptivity, yet it disregards the interchannel relationship that regularly corresponds to various objects' feature maps. Large convolution kernels provide a large and effective receptive field [14]. The optical input contains abundant feature information. Merging with depthwise [15] model design techniques, we use a large kernel attention to attain long-range dependencies between various channels with the optical input, which is composed of depthwise convolution (K^{DW}), depthwise dilation convolution ($\mathrm{K}^{\mathrm{DW\_D}}$) and 1×1 convolution ($\mathrm{K}^{\mathbf{1\times1}}$). MMF can be written

as the following:

$$\mathbf{G}_{k,l,m} = \sum_{i,j} \mathbf{K}^{\mathbf{DW}}{}_{i,j,m} \cdot \mathbf{F}^{\mathbf{RGB}}{}_{k+i-1,l+j-1,m} \tag{1}$$

$$\hat{\mathbf{G}}_{k,l,m} = \sum_{i,j} \mathbf{K}^{\mathbf{DW\_D}}{}_{i,j,m} \cdot \mathbf{G}_{k+i-1,l+j-1,m} \tag{2}$$

$$\hat{\mathbf{G}}_{k,l,n} = \sum_{i,j} \mathbf{K}^{\mathbf{1\times 1}}{}_{i,j,m,n} \cdot \hat{\mathbf{G}}_{k+i-1,l+j-1,m} \tag{3}$$

$$\mathbf{Output} = \hat{\mathbf{G}}_{k,l,n} \otimes \mathbf{F}^{\mathbf{DSM}} \tag{4}$$

F ($\mathrm{F} \in R^{C\times H\times W}$) and G are the feature maps. *m, n* indicate the number of input and output channels. Channel numbers of **F** and **G** are unified after the depthwise convolution. As a result of the relatively limited feature information in DSM, we only extract coarse object contextual representations from this input. Depthwise convolution aims to exploit the relationship between spatial and local areas, whereas dilation convolution is for enlarging the receptive field. $\otimes$ denotes the elementwise product where $\mathrm{G}_{\boldsymbol{k,l,n}}$ acts as the attention weight for the fusion with DSM input. Therefore, we could explore the multimodality complementary context from both inputs.

C. Decoder

We use MLP layers as the decoder, which eliminates the need to design complex components for feature extraction and greatly reduces memory consumption [9]. After MLP and upsampling, all the output feature maps $\mathcal{X}$ ($\mathcal{X}_i \in \mathbb{R}^{(W/2^i)\times(H/2^i)\times C_i}$) from the four branches in HRNet are reduced to 1/4 of the original size, with the identical channel number. Equation (5) can represent this process, where Z_i is the output

$$Z_i = UP_{(\frac{W}{4},\frac{H}{4})}(\mathrm{MLP}_{v1}X_i) \tag{5}$$

$$Z_2 = \mathrm{Conv}_{\mathrm{cls}}(\mathrm{MLP}_{v2}(\mathrm{Concat}(Z_i))). \tag{6}$$

Following the concatenation of Z_i, the fused representation is obtained via MLP_{v2}. Finally, the pixel classification is obtained through 1×1 convolution ($\mathrm{Conv}_{\mathrm{cls}}$). Equation (6) shows the process for acquiring segmentation mask.

III. Experiments and Analysis

A. Dataset

The aerial datasets in the experiment originate from Potsdam and Vaihingen provided by ISPRS,[1] which have an average resolution of 6000×6000 pixels and 2500×2500 pixels. They consist of cars, low vegetation, buildings, trees, impervious surfaces, and cluttered backgrounds. Combining with DSM, red, green and blue (RGB) and near infrared, red and green (IRRG) channel compositions are selected for Potsdam and Vaihingen datasets, respectively. Both are split into the training set, validation set and test set according to the ratio of 0.8, 0.1, and 0.1. We conduct an exploratory analysis and find that cars account for the least, less than 1.7%. In Vaihingen, 8.57% of trees are more prevalent than in Potsdam and other species proportions are comparable. Constrained by the GPU memory, we crop the original high-resolution image into 512×512 and also perform data augmentation, including adjusting image brightness, contrast, rotation, symmetric padding, etc., to reduce overfitting during the training process. As can be noted in Fig. 3, both datasets suffer from the class imbalance issue. Impervious surfaces are approximately 16 times more numerous than cars. Consequently, there will be a significant bias toward head classes, resulting in the decreased tail class performance. RSIs in ISPRS are captured in normal conditions. To verify that our model is robust in foggy weather, motivated by the robustness evaluation approach in [16], we rendered five different levels of fog for each image in the test set using diamond-square algorithm. Detailed results are shown in Fig. 4.

[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*).

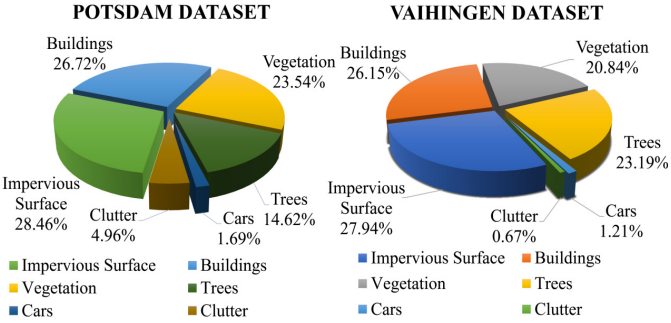

Fig. 3. Class imbalance exists in both datasets.

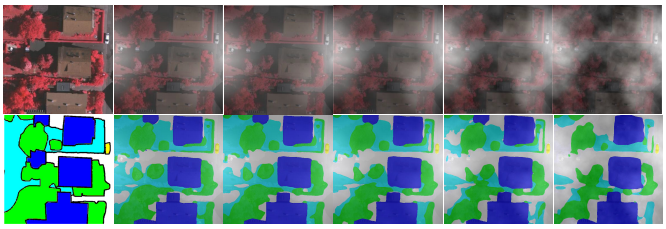

Fig. 4. Renderings of five levels of fog severity using Vaihingen dataset. First column: clean and ground truth (GT), others: fog of different severity levels with corresponding land cover classification results.

B. Metrics

The evaluation metrics in the experiments are overall accuracy (OA), mean Intersection over Union (mIoU), and corruption degradation (CD). CD, derived from [16], is the average of degradations across various severity levels which is utilized to evaluate the robustness. Definition of CD is given below

$$\mathrm{D} = 1 - \mathrm{IoU} \tag{7}$$

$$\mathrm{CD}^{\mathrm{f}} = \frac{\widetilde{\mathrm{D}^{\mathrm{f}}}}{\widetilde{D^{\mathrm{ref}}}} \times 100\%. \tag{8}$$

Higher **mIoU** and **OA** values are associated with more accurate land cover classifications. **ref** and **f** are the reference and selected model correspondingly. **D** represents the degradation where $\widetilde{\mathrm{D}^{\mathrm{f}}}$ stands for the mean degradation acquired across different severity levels of fog corrupted dataset. $\mathbf{CD}^{\mathbf{f}}$ evaluates the absolute robustness performance of method **f**

TABLE I

ABLATION STUDY ON THE VARIANTS OF VAIHINGEN TEST SET

Method	Imp Surf *			Building			Low Veg *			Tree			Car			mIoU(%)			OA(%)		
	Clean↑	Fog↑	CD↓	Clean↑	Fog↑	CD↓	Clean↑	Fog↑	CD↓	Clean↑	Fog↑	CD↓	Clean↑	Fog↑	CD↓	Clean↑	Fog↑	mCD↓	Clean↑	Fog↑	CD↓
A: w/o DSM	84.30	67.18	106.59	87.79	73.29	107.92	67.76	54.79	105.19	76.34	65.26	109.35	68.99	49.15	104.80	77.04	61.93	106.77	85.07	77.87	105.73
B: w/o IRRG	74.17	56.06	**142.71**	78.55	64.95	**141.62**	54.92	48.88	**118.94**	66.56	55.76	**139.25**	58.16	34.24	**135.53**	66.47	51.98	**135.61**	75.46	67.13	**157.05**
C: MCC	84.31	68.42	102.57	88.55	74.35	103.64	71.32	56.93	100.21	**78.92**	67.10	103.56	70.46	50.63	101.75	78.71	63.49	102.34	87.16	78.14	104.44
D: w/o MMF	83.52	66.32	109.39	86.63	71.26	116.12	67.65	52.73	109.98	75.05	64.90	110.48	67.90	41.63	120.30	76.15	59.37	113.25	85.44	77.02	109.79
E: Ours	**86.36**	**69.21**	100.00	**89.92**	**75.25**	100.00	**72.02**	**57.02**	100.00	78.10	**68.23**	100.00	**70.30**	**51.48**	100.00	**79.34**	**64.24**	100.00	**88.62**	**79.07**	100.00

TABLE II

COMPARISON STUDY ON THE VARIANTS OF POTSDAM TEST SET

Method	Imp Surf *			Building			Low Veg *			Tree			Car			mIoU(%)			OA(%)		
	Clean↑	Fog↑	CD↓	Clean↑	Fog↑	CD↓	Clean↑	Fog↑	CD↓	Clean↑	Fog↑	CD↓	Clean↑	Fog↑	CD↓	Clean↑	Fog↑	mCD↓	Clean↑	Fog↑	CD↓
FCN [5]	70.08	25.64	**136.87**	71.06	37.63	**243.54**	68.41	6.10	**105.73**	66.72	14.24	**167.66**	41.46	13.20	**262.95**	63.55	19.36	**183.35**	78.41	40.76	**155.04**
RefineNet [18]	75.15	36.01	117.78	77.94	55.52	173.68	70.07	8.76	102.74	69.54	36.22	124.69	61.96	42.03	175.61	70.93	35.71	138.90	82.07	46.01	141.30
DeepLabV3+ [19]	76.65	37.48	115.07	78.76	64.50	138.62	72.55	9.36	102.06	69.41	40.14	117.03	58.24	49.56	152.80	71.12	40.21	125.12	82.39	49.84	131.27
PSPNet [20]	78.97	35.85	118.07	76.67	61.44	150.57	73.71	9.26	102.17	68.70	38.16	120.90	61.73	47.08	160.32	71.96	38.36	130.41	80.32	47.26	138.03
SETR [21]	76.45	37.44	115.15	75.44	62.03	148.26	70.65	10.20	101.11	69.69	40.49	116.34	62.16	52.58	143.65	70.88	40.55	124.90	85.03	50.06	130.70
OCRNet [22]	78.24	44.21	102.69	85.50	69.81	117.88	74.67	**11.41**	99.75	73.49	46.12	105.34	76.14	63.22	111.42	77.61	46.95	107.42	87.67	58.50	108.61
Segmenter [23]	77.77	40.50	109.52	82.88	55.09	175.36	71.76	9.24	102.20	72.92	40.36	116.60	74.25	51.45	147.08	75.92	39.33	130.15	83.70	49.45	132.30
SegFormer [9]	78.85	42.81	105.26	84.33	70.73	114.29	72.72	11.03	100.18	72.89	45.41	106.73	75.96	62.55	113.45	76.95	46.51	107.98	87.43	56.22	114.58
Ours	**84.20**	**45.67**	100.00	**89.92**	**74.39**	100.00	**74.54**	11.19	100.00	**77.94**	**48.85**	100.00	**80.19**	**66.99**	100.00	**81.36**	**49.42**	100.00	**89.25**	**61.79**	100.00

TABLE III

VARIOUS BACKBONES ON CORRUPTED VAIHINGEN TEST SET

Backbone	mIoU(%)↑	OA(%)↑	CD(%)↓
ResNet-101	51.45	70.88	135.77
ResNext-101	57.64	72.06	118.46
MiT-B5	61.28	77.10	108.28
MobileNetV2	47.10	65.85	**147.93**
HRNet-W48	**64.24**	**79.07**	100.00

and the higher portion above 100% indicates a degradation compared with the reference model in terms of robustness.

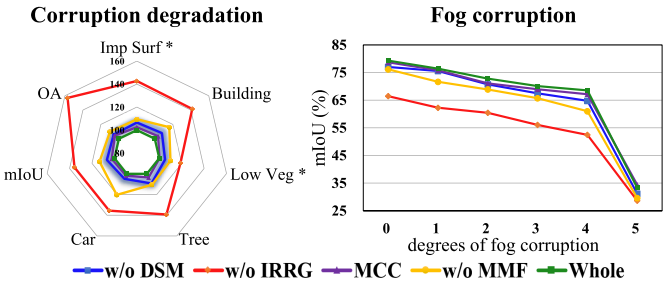

Fig. 5. Ablation study visualization on fog corrupted Vaihingen test set. In the radar plot, a smaller envelop indicates better robustness. Increasing fog severity shows mIoU degradation.

C. Implementation Details

Image with the ID "4-12" in the Potsdam dataset has been removed due to the label error. All the experiments were conducted on two Titan XP with SyncBN and LayerNorm. We adopt AdamW optimizer and initial learning rate is set to 0.00006. Poly optimization strategy is adopted with the power and weight decay setting as 1 and 0.01, respectively. Batch size is 10 and we train the model for 288 epochs. For handling class imbalance existing in both datasets, we employ *unified focal loss* [17] function for the robustness and training convergence.

D. Results and Analysis

Clean and *Fog* in Tables I and II represent the clean and fog corrupted test set, respectively. All the columns except *OA* are *IoU*. The **higher** the *IoU* of *Clean* and *Fog*, the more accurate the land cover classification will be. Models with **lower** *CD* perform better in foggy conditions and are more robust than the reference model.

1) Ablation Study: To verify the effectiveness of the backbone, an ablation study was conducted using fog corrupted test set. The backbone in Fig. 2 was replaced with ResNet-101, ResNext-101, MiT-B5 [9], MobileNetV2 [15], HRNet-W48 [7]. As can be concluded from Table III, MobileNetV2 has inferior classification accuracy and robustness in view of a small number of parameters. Mit-B5 shows a relatively better performance in robustness when fog exists. HRNet-W48 demonstrates a 2.96% improvement in mIoU along with an 8.28% reduction in CD over MiT-B5.

To validate the effectiveness of each element, we conduct an ablation study on the Vaihingen dataset and results are shown in Table I and Fig. 5. **w/o** in Table I represents **without**, w/o DSM along with w/o IRRG represent we only use the data from the same modality with identity mapping, thus verifying that model can learn the complementary relationship from multimodal inputs to enhance the robustness. **MCC** represents replacing MMF with multichannel concatenation, thereby testing whether pure channel stacking or MMF is better at fusing multimodal data. w/o MMF indicates replacing MMF with 1×1 convolution and sum operations. Our model is set as the reference. From the comparison of A and B in Table I, it is worth noting that mIoU performance on the clean data of DSM has decreased by 10.57% due to the lack of IRRG. Mean CD (mCD), which increases by 28.84%, demonstrates the degradation of robustness without IRRG during training. Looking into the results A, B, and C, multimodality outperforms the single one in mIoU by nearly 2%. Observing C and E, MMF will bring a decrease in mCD of 2.34%. The comparison between D and E indicates that MMF is more effective at obtaining the complementary characteristics of multimodality than 1×1 convolution. There will be an almost 10% and 3% improvement in robustness and accuracy, respectively.

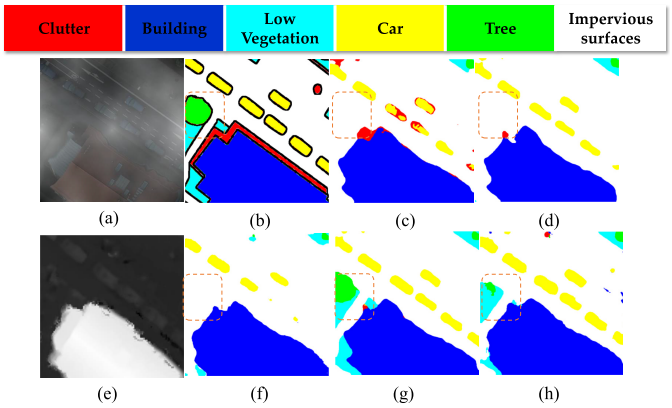

Fig. 6. Qualitative comparisons of several approaches for semantic segmentation of RSIs. The original image is corrupted by the fog of severity level three. (a) Corrupted image. (b) GT. (c) HRNet. (d) W/o MMF. (e) DSM. (f) Segformer. (g) Ours on clean. (h) Ours on fog.

Fig. 5 (right) demonstrates the performance on the corrupted Vaihingen dataset with different severity levels of fog. *Level0* denotes the clean dataset. Model trained only on the DSM input has the worst robustness. The difference between each method is not significant when the corruption degree is within 1–3. Models with MMF and MCC are characterized with better performance in both accuracy and robustness.

2) Comparison Study: To compare the performance of different methods, we conducted a comparison study on the Potsdam test set. The selected methods can be separated into two groups: CNN-based models whose backbone is ResNet101 (including FCN [5], RefineNet [18], DeepLabV3+ [19], PSPNet [20], OCRNet [22]) and Transformer-based models (like SETR [21], Segmenter [23], and Segformer [9] using DeiT-B, DeiT-B and MiT-B5 [9] as the backbone). The number of parameters is approximately 70–90 M for each model. In all cases except ours, the input multimodal data are concatenated over one dimension. It can be observed from Table II that OCRNet perform best excluding ours. Our model, compared to OCRNet, delivers a 3.75% and 2.47% improvement in mIoU performance on *Clean* and *Fog*, respectively, as well as a 7.42% reduction in mCD. In terms of robustness to natural corruption noises, Transformers perform worse than CNNs due to the requirement of large amount of data for pretraining. The inference result on corrupted image under severity level 3 is illustrated in Fig. 6(a). Model that is sensitive to natural noise, such as Fig. 6(c), (d), and (f), cannot classify trees and low vegetation accurately as a result of dense fog in the box area. Ours can well capture the long-range dependencies, not only can precisely label the objects in the box area, but also facilitate fine-grained classification of the edges. Our model shows superior robustness and an improvement in the land cover classification of fog corrupted scenes comparing to previous ones.

IV. Conclusion

In this letter, we propose an end-to-end robust framework for the classification of land covers. Based on HRNet, we show that complementary features from multimodality fusion can be learned by MMF to improve robustness performance. Experiment results indicate that our model generalizes well on the fog corrupted test set and achieves a balance between accuracy and robustness. We hope that this letter will motivate scholars and practitioners in this area to further investigate how to design a robust model in the remote sensing domain.

References

[1] S. Lorenz, P. Ghamisi, M. Kirsch, R. Jackisch, B. Rasti, and R. Gloaguen, "Feature extraction for hyperspectral mineral domain mapping: A test of conventional and innovative methods," *Remote Sens. Environ.*, vol. 252, Jan. 2021, Art. no. 112129.

[2] X. Zhang, W. Ma, C. Li, J. Wu, X. Tang, and L. Jiao, "Fully convolutional network-based ensemble method for road extraction from aerial images," *IEEE Geosci. Remote Sens. Lett.*, vol. 17, no. 10, pp. 1777–1781, Oct. 2020.

[3] S. Kunwar *et al.*, "Large-scale semantic 3-D reconstruction: Outcome of the 2019 IEEE GRSS data fusion contest—Part A," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 922–935, 2021.

[4] H. Hichri, Y. Bazi, N. Alajlan, and S. Malek, "Interactive segmentation for change detection in multispectral remote-sensing images," *IEEE Geosci. Remote Sens. Lett.*, vol. 10, no. 2, pp. 298–302, Mar. 2013.

[5] 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, Apr. 2017.

[6] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," 2015, *arXiv:1505.04597*.

[7] J. Wang *et al.*, "Deep high-resolution representation learning for visual recognition," 2019, *arXiv:1908.07919*.

[8] S. Zheng *et al.*, "Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers," in *Proc. CVPR*, Jun. 2021, pp. 6877–6886.

[9] E. Xie, W. Wang, Z. Yu, A. Anandkumar, J. M. Alvarez, and P. Luo, "SegFormer: Simple and efficient design for semantic segmentation with transformers," 2021, *arXiv:2105.15203*.

[10] X. Zheng, T. Gong, X. Li, and X. Lu, "Generalized scene classification from small-scale datasets with multitask learning," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–11, 2022.

[11] H. Ning, B. Zhao, and Y. Yuan, "Semantics-consistent representation learning for remote sensing image–voice retrieval," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–14, 2022.

[12] X. Zheng, Y. Yuan, and X. Lu, "A deep scene representation for aerial scene classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 7, pp. 4799–4809, Jul. 2019.

[13] Y. Cao, J. Xu, S. Lin, F. Wei, and H. Hu, "GCNet: Non-local networks meet squeeze-excitation networks and beyond," 2019, *arXiv:1904.11492*.

[14] M.-H. Guo, C.-Z. Lu, Z.-N. Liu, M.-M. Cheng, and S.-M. Hu, "Visual attention network," 2022, *arXiv:2202.09741*.

[15] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "MobileNetV2: Inverted residuals and linear bottlenecks," Mar. 2019, *arXiv:1801.04381*.

[16] 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.

[17] M. Yeung, E. Sala, C.-B. Schönlieb, and L. Rundo, "Unified focal loss: Generalising dice and cross entropy-based losses to handle class imbalanced medical image segmentation," *Comput. Med. Imag. Graph.*, vol. 95, Jan. 2022, Art. no. 102026.

[18] G. Lin, A. Milan, C. Shen, and I. Reid, "RefineNet: Multi-path refinement networks for high-resolution semantic segmentation," in *Proc. CVPR*, Jul. 2017, pp. 5168–5177.

[19] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," 2018, *arXiv:1802.02611*.

[20] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," Apr. 2017, *arXiv:1612.01105*.

[21] S. Zheng *et al.*, "Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers," Jul. 2021, *arXiv:2012.15840*.

[22] Y. Yuan, X. Chen, X. Chen, and J. Wang, "Segmentation transformer: Object-contextual representations for semantic segmentation," Apr. 2021, *arXiv:1909.11065*.

[23] R. Strudel, R. Garcia, I. Laptev, and C. Schmid, "Segmenter: Transformer for semantic segmentation," Sep. 2021, *arXiv:2105.05633*.